Post-Genomic Approaches to Personalized Medicine: Applications in Exome Sequencing, Microbiome, and COPD

Dissertation Advisor: Professor John Quackenbush Jarupon Fah Sathirapongsasuti

Post-Genomic Approaches to Personalized Medicine:

Applications in Exome Sequencing, Microbiome, and COPD

Abstract

Since the completion of the sequencing of the human genome at the turn of the century, genomics has revolutionized the study of biology and medicine by providing high-throughput and quantitative methods for measuring molecular activities. Microarray and next generation sequencing emerged as important inflection points where the rate of data generation skyrocketed. The high dimensionality nature and the rapid growth in the volume of data precipitated a unique computational challenge in massive data analysis and interpretation. Noise and signal structure in the data varies significantly across types of data and technologies; thus, the context of the data generation process itself plays an important role in detecting key and oftentimes subtle signals. In this dissertation, we discuss four areas where contextualizing the data aids discoveries of disease-causing variants, complex relationships in the human microecology, interplay between gene and environment, and genetic regulation of gene expression. These studies, each in its own unique way, have helped made possible discoveries and expanded the horizon of our understanding of the human body, in health and disease.

Table of Contents

Acknowledgements

I would like to express my deepest appreciation to my advisor Professor John Quackenbush, whose aptitude and attitude toward scholarship and life I aspire to emulate. From our first phone conversation before I matriculated in the program to our most recent email exchanges, he never fails to convey a sense of excitement and commitment to science. Without his support, guidance, and patience, this dissertation would not have been realized.

I would like to thank my committee members for guidance throughout my PhD career. Professor Curtis Huttenhower, my first-year academic advisor, showed me an example of a highly accomplished young scientist through whom I had the opportunity to be a part of the Human Microbiome Project, one of the highlights of my PhD career. I am grateful to Professor Dawn DeMeo for the countless hours she spent grooming the LGRC dataset, editing my manuscript, and guiding me through my last two years. Finally, I am thankful for the opportunity to work with Professor Christoph Lange on the last chapter of my dissertation.

I would like to acknowledge the wonderful support network of friends and colleagues. Members of Department of Biostatistics, Quackenbush Lab, and Dana-Farber's Center for Cancer Computational Biology are among the most talented and kind individuals I've had the privilege to work with. A particular thanks go to Dr. Aedin Culhane, Dr. Lan Hu, Dr. Yaoyu Wang, Dr. John Platig, Ms. Julianna Coraccio, Ms. Joan Coraccio, and Ms. Jelena Follweiler for their friendship and support over the years.

I thank members of Adams House for a home and a family. I could not overstate how thankful I am for the opportunity to be a part of the community and to help contribute to its vitality socially and intellectually.

Thank you, Gabe Malseptic, Emmy Smith, Amos Irwin, Aedin Culhane, Simon French, and Tanmoy Laskar for lending your places during my last few months in Boston. My brief time with each of you was a sweet memory I will forever cherish.

Thank you, Mio Sakata, for standing beside me through this wondrous journey. Without you I would not have gotten through the tough times. And without you I would not have had anyone to share the good time and the accomplishment with.

Finally, a big "thank you" to my loving family: my sister Som, mother Neilvadee, and father Dr. Kreingkrai Sathirapongsasuti. Your love and trust warm my heart and inspire me to not only be a better scientist but a better person, for Thailand and for the world. I hope I have made you proud. This PhD is for you.

List of Figures

Chapter 1: Introduction

Background

The past decade has seen the power of genomics for biomedical research. The greatest impact of genomics has been the ability to study biological system in a comprehensive, unbiased, hypothesis-free manner. In the study of disease, genomic technologies enable a systematic approach to discover the genes and cellular pathways underlying disease. To date, the genomic approach has resulted in the identification of over 2,850 genes underlying Mendelian diseases, over 1,100 loci affecting common disorders, and hundreds of carcinogenic somatic mutations [1].

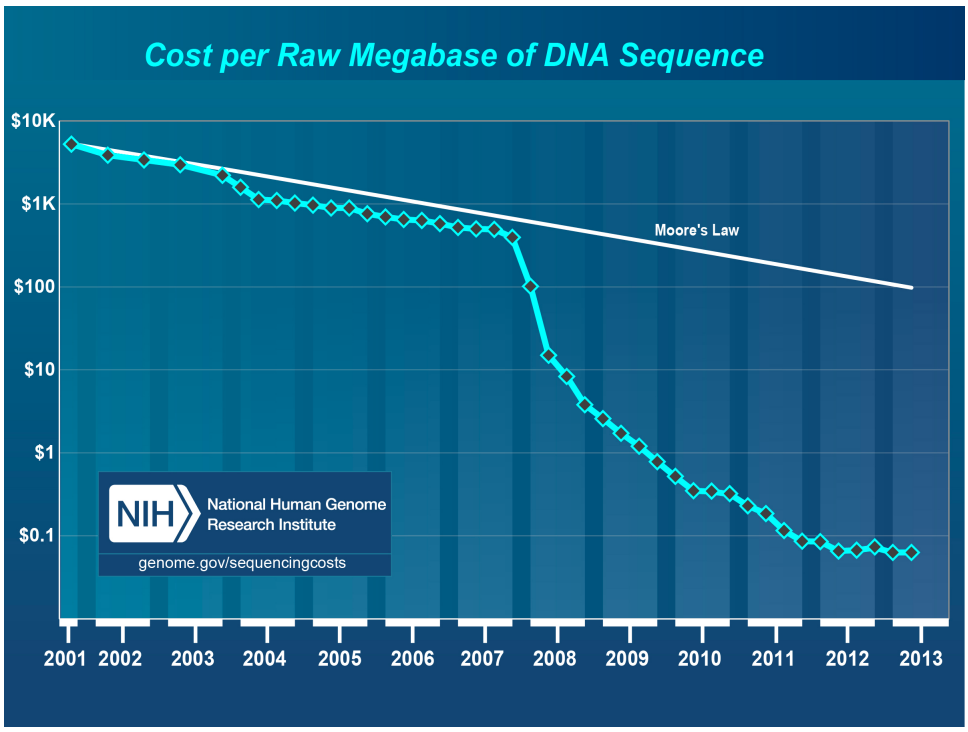
The high-throughput nature of genomic technologies has enabled large-scale interrogations of the biology of genome and of disease. Several genetic epidemiological studies now include hundreds, if not thousands, of subjects. The speed and cost of sequencing technologies has been a key driver of this revolution. As shown in Figure 1.1, the cost per base and per genome of sequencing has dramatically and steadily dropped from over a hundred million dollars to a few thousand dollars[1]. As a result, the amount of genomic data produced has exponentiated to a point where the bottleneck is analysis and interpretation—a problem sometimes called "the $1,000 genome, the

[1] http://www.genome.gov/sequencingcosts/

$100,000 analysis [2]." Thus efficient computational and statistical analysis methods are critical in harnessing the full power and promise of genomic technologies.

Through a common platform of sequencing technologies, many types of genomic information can be generated, for example gene expression, epigenetic markings, DNA point mutations, structural variations, and microbial compositions. With the richness of data types comes complexity. Although generated by a common sequencing platform, analyzing and interpreting these data often require specialized methods. This is primarily because each type of genomic data possesses unique signal and noise characteristics. In this dissertation, we explored unique characteristic of several types of genomics data and developed approaches to extract information and understanding from them. The four application areas are: exome sequencing-based copy number variation detection, microbial co-occurrence network in the human microbiome, sexual dimorphism in chronic obstructive pulmonary disease (COPD), and bicluster-based error rate control procedure for expression quantitative trait loci (eQTL) analysis.

A.



B.

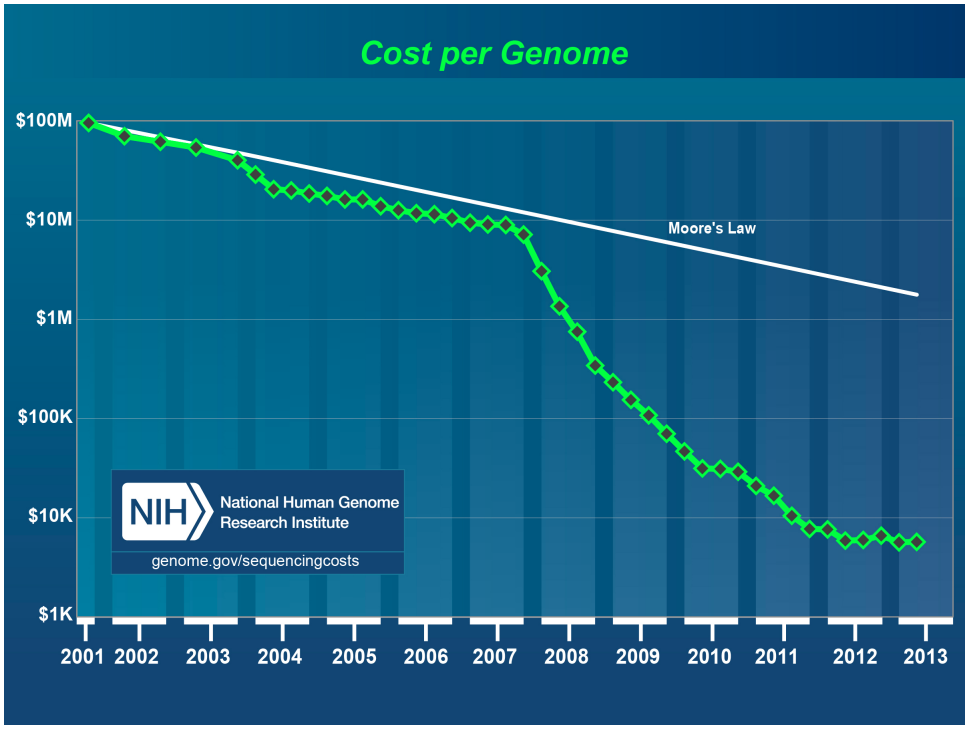


Figure 1.1: Reduction in sequencing cost (A) per megabase and (B) per genome. Technology developments that keep up with Moore's law are considered to be doing exceedingly well. The sudden out-pacing in 2008 corresponds to the transition from Sanger sequencing to next-generation sequencing technologies. Source: www.genome.gov/sequencingcosts/

ExomeCNV: Exome Sequencing-Based Copy Number Variation and Loss of Heterozygosity Detection

Exome sequencing is a strategy to selectively sequence the coding regions of the genome as a low-cost yet effective alternative to whole genome sequencing. We are interested in detecting structural variation, which is a class of large-scale mutation commonly found in cancer genomes. The target enrichment step in the library preparation process substantially biases the sequencing information. In addition to the technical noise, heterogeneity of the sample introduces an additional layer of bias. In Chapter 2, we detail the problems and developed ExomeCNV, a structural variation detection method that accounts for characteristic biases in exome sequencing data.

Co-Occurrence Relationships in the Human Microbiome

The interplay between microbes and human health has recently gained attention in the biomedical field. The advancement in sequencing technology allows us to interrogate microbial composition of micro-communities in the context of human hosts. We are interested in studying the patterns of co-occurrence and co-exclusion between microbes. Assessing correlation between such compositional data turned out to be quite challenging. As the individual components in a compositional data have to add up to one, spurious correlation may arise as a result of normalization. We developed a method called ReBoot, which mitigated this compositional effect and aided discovery of microbial interactions in healthy human microbiome. We discussed in Chapter 3 the challenges and methods that might help overcome the spurious correlation problem.

Integrative Genomics of Sexual Dimorphism in COPD

Chronic obstructive pulmonary disease (COPD) is the third leading cause of death in the US. The number of women dying of the disease has been increasing drastically over the last twenty years and now surpasses that of men. An cumulating body of evidence suggests that women may be biologically more susceptible to COPD. However, no study has done a systematic screen of genes or pathways that are differentially expressed between genders. Using a unique set of data from Lung Genomics Research Consortium, we developed the sexual dimorphic and COPD differential (SDCD) analysis, a stratification analysis that highlights the elusive sexual dimorphic effects in gene expression data. Chapter 4 discusses the methods and findings of this work.

BBER: Bicluster-Based Error Rate Control for eQTL Analysis

The availability of multiple genomic data types enables integrative approaches that reveal correlation information across data types. Correlating genotypes and gene expression data, or expression quantitative trait loci (eQTL) analysis, shed light on the genetic regulation of gene expression. Because of the ultra-high dimensional nature of eQTL analysis, multiple hypotheses testing adjustment is very critical in keeping the number of false positives at a reasonable rate. Traditional approaches such as Bonferroni or Benjamini-Hochberg procedures assume independence between tests, but in reality genetic markers are highly correlated and the eQTL tests are rarely independent. In Chapter 5, we discussed the development of BBER a technique to cluster correlated eQTL tests to improve sensitivity of the eQTL analysis.

Chapter 1 Bibliography

1. Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470: 187-197.

2. Mardis ER (2010) The $1,000 genome, the $100,000 analysis? Genome Med 2: 84.

Chapter 2: Exome Sequencing-Based Copy-Number Variation and Loss of Heterozygosity Detection: ExomeCNV

Abstract

The ability to detect copy-number variation (CNV) and loss of heterozygosity (LOH) from exome sequencing data extends the utility of this powerful approach that has mainly been used for point or small insertion/deletion detection. We present ExomeCNV, a statistical method to detect CNV and LOH using depth-of-coverage and B-allele frequencies, from mapped short sequence reads, and we assess both the method's power and the effects of confounding variables. We apply our method to a cancer exome resequencing dataset. As expected, accuracy and resolution are dependent on depth-of-coverage and capture probe design.

Introduction

The development of next-generation sequencing has enabled routine large-scale resequencing projects, permitting us to perform increasingly more comprehensive DNA variant analysis. However, the cost and analytical complexity of sequencing still limit the number of whole genomes that can be sequenced in any single project [1]. In fact, the analysis of complete human genome sequence often interprets DNA alterations in protein coding regions primarily. This is in practice a reasonable strategy since approximately 85% of the disease-causing mutations are found in the coding regions or canonical splice sites [2]. Thus, whole-exome sequencing presents an effective alternative to whole-genome sequencing and provides an unbiased, cost-effective, and time-efficient tool for the study of the genetic basis for disease.

Following the first successful application of whole-exome sequencing in re-discovering the cause of a dominantly inherited rare Mendelian disorder Freeman-Sheldon syndrome [3], a number of studies have reported similar successes [4-9]. Although the cost of both genome and exome sequencing continues to fall at a rapid pace, whole-exome sequencing has a number of advantages, including a lower cost, more straight-forward data analysis and interpretation, and significantly greater depth of coverage with a corresponding overall improvement in data quality. Exome sequencing is rapidly becoming a fundamental tool for genetic and functional genomic research laboratories and a diagnostic tool in clinics. At present the main applications of targeted exonic sequencing is for the determination of single nucleotide variants (SNVs) or small indel variants but not structural variation.

Structural variation, especially copy-number variation (CNV) or loss of heterozygosity (LOH), is an important class of genetic variability in Mendelian, common inherited diseases, and cancer [10-17]. As is true of SNVs, there are population-specific, common CNVs and rare, disease-causing CNVs [18,19]. Many large-scale projects [20-22] and technological platforms [23,24] have been devised to estimate the prevalence and impact of CNV. Array Comparative Genomic Hybridization (CGH)[23] and SNP genotyping arrays have been widely used as standard methods to detect CNV and LOH. However, with the rapid increase in genomic and exomic sequence, there is growing interest in the use of these data to detect CNVs.

While methods have been developed for CNV estimation in whole-genome sequencing [25,26], these methods make key assumptions that fail to hold in the exome sequencing setting. For example, Yoon et al. [26] assumes random, unbiased distribution of sequence reads, such

that read depth can be modeled as a normal distribution across the genome, and deviation from the background indicates the presence of CNV. This random read distribution assumption breaks down in the context of exome capture as the probes have variable specificity and efficiency for the targeted exonic regions. The discrete nature of exome sequences also presents problems to existing methods. Many whole-genome CNV detection tools use segmentation algorithms that assume continuity of search space and do not function properly when given discontinuous and variable length exome sequencing data. SegSeq [25], for example, merges windows of a fixed length based on a log-ratio difference statistic. Lastly, because exons are generally smaller than insert sizes for paired-end sequencing (200-500 bp), paired-end based CNV detection methods are not generally applicable to exome data.

Here we present ExomeCNV, which uses depth-of-coverage and B-allele frequencies from mapped short sequence reads to estimate copy number variation (Figure 2.1, left side) and loss of heterozygosity (Figure 2.1, right side). We describe an assessment of its validity, sensitivity, specificity, and limitations through an analysis of a melanoma tumor and a matched normal sample. Important model assumption and the effect of important confounding factor such as sample admixture rate are also considered.

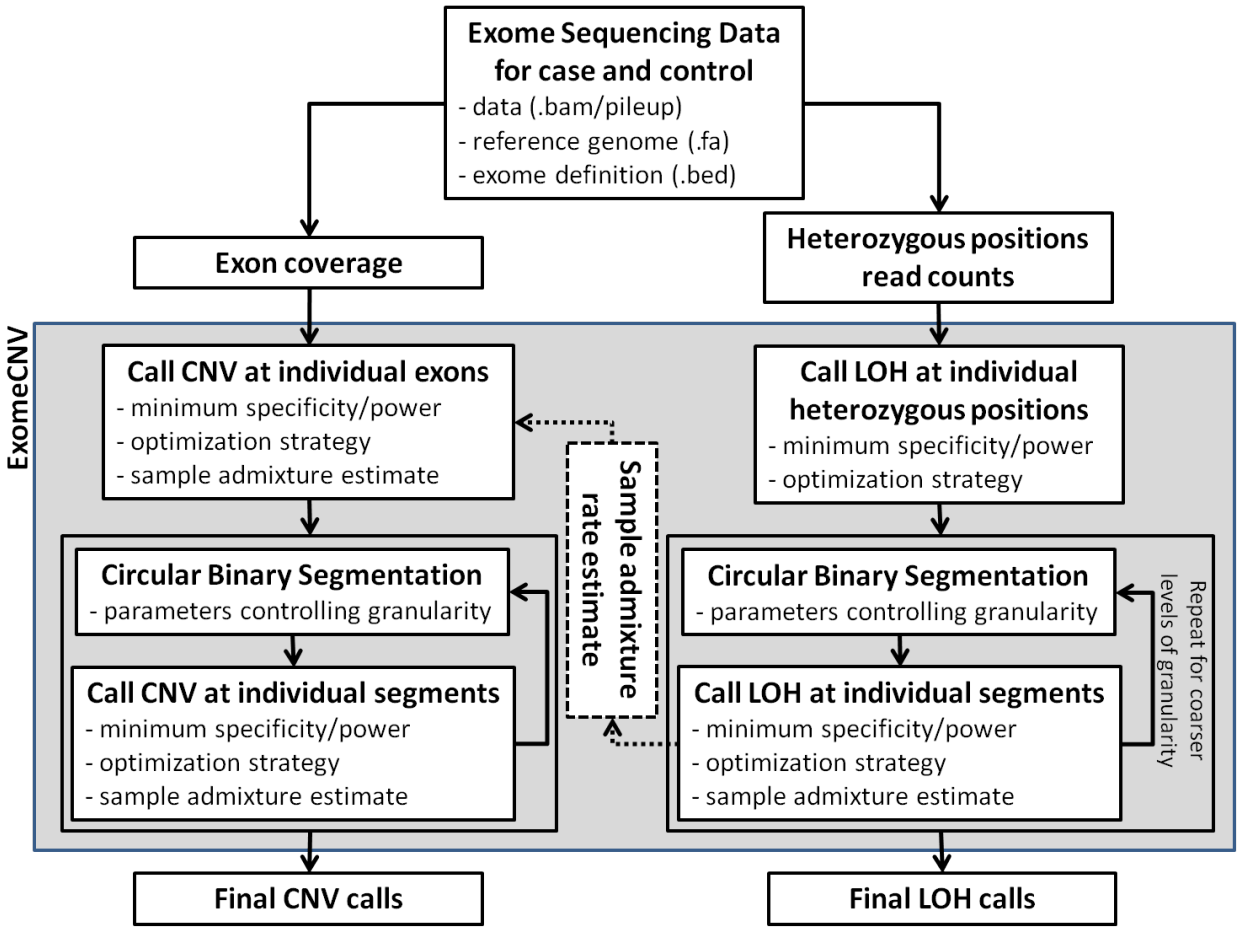


Figure 2.1 Overview of ExomeCNV analysis workflows. Two workflows are present: CNV detection and LOH detection. Each involves similar steps of exon/position/segment-wise CNV/LOH calling, Circular Binary Segmentation, and interval merging. User inputs and parameters are listed at each step.

Methods

Correlation of Depth-of-Coverage

To plot the correlation of depth-of-coverage, we used the internal exome data that were captured by the same Agilent SureSelect Human All Exon Kit and sequenced on the Illumina GAIIx. Samples 1 through 4 were generated by 2 lanes of 76bp single-end sequencing, sample 5 was generated by 3 lanes of 76bp single-end sequencing and sample 6 was generated by 1 lane of 76+76bp paired-end sequencing. For each sample, the average depth-of-coverage per exon

was normalized by dividing the average coverage by the overall exome average coverage and then, the normalized depth-of-coverage were compared between 15 pairs of samples.

The CNV and LOH Detection Algorithm

Power analysis of CNV Detection

Consider an exon of length L, let X and Y denote the numbers of reads, each of length w, mapped within the exon in question in case (e.g. tumor) and control (e.g. matched normal), respectively. The depth-of-coverage is then Xw/L and Yw/L for case and control, respectively. Although we discuss our method in terms of depth-of-coverage, our method is developed in terms of the count statistics X and Y. Let N_X and N_Y be the total numbers of aligned reads in case and control, respectively. Define the read count ratio:

$$R = \frac{X/N_X}{Y/N_Y}.$$

We divide the raw counts X and Y by the total number of reads N_X and N_Y to mitigate the effect of overall increase in local counts due to the increase in total depth-of-coverage. Finally, we adjust the ratio so that the exome-wide median is 1. Without lost of generality, we assume $N_X = N_Y$ and reduce $R = X/Y$. Because X and Y follow Poisson distributions with parameters λ_X and λ_Y, respectively, with sufficient depth-of-coverage the Poisson distributions converge to a normal distribution with equal means and variances: $N(\lambda_X, \lambda_X)$ and $N(\lambda_Y, \lambda_Y)$. Under the null hypothesis of no CNV, $\lambda_X = \lambda_Y$, and under the alternative hypothesis, $\lambda_X = \rho\lambda_Y = \rho\lambda$. ρ indicates the copy-number ratio; for example, $\rho = 0.5$ for deletion, and $\rho = 1.5$ for duplication. By Geary-Hinkley transformation [27-29], let

$$t(\rho) = \frac{\mu_Y R - \mu_X}{\sqrt{\sigma_Y^2 R^2 + \sigma_X^2}} = \frac{\lambda_Y R - \lambda_X}{\sqrt{\lambda_Y R^2 + \lambda_X}} = \frac{\lambda R - \rho\lambda}{\sqrt{\lambda R^2 + \rho\lambda}} = \frac{(R-\rho)\sqrt{\lambda}}{\sqrt{R^2+\rho}},$$

and $t(\rho)$ follows the standard normal distribution. Thus, the specificity and sensitivity are $1 - \alpha$ and $1 - \beta$ where

$$\alpha = \begin{cases} \Phi\big(t(1)\big) & \text{if } \rho < 1, \\ 1 - \Phi\big(t(1)\big) & \text{if } \rho \geq 1 \end{cases}$$

$$\beta = \begin{cases} 1 - \Phi\big(t(\rho)\big) & \text{if } \rho < 1, \\ \Phi\big(t(\rho)\big) & \text{if } \rho \geq 1 \end{cases}$$

The formulas above describe the achievable specificity and sensitivity of a given cutoff ratio R. Considering values of R from zero to infinity, we can plot the receiver operating characteristic (ROC) curve (Appendix 2). In practice, we may wish to identify a cutoff $r(\rho)$ which yields the desired minimum specificity and/or sensitivity for testing a particular copy-number ratio ρ at an exon with certain depth-of-coverage and length. This can be achieved by solving above equations, and an appropriate cutoff $r(\rho)$ can be chosen from a set of solutions to maximize user-selected quantity metrics such as specificity, sensitivity, or area under curve.

In the presence of sample admixture, the "true" copy-number ratio will tend to 1. In particular, if a fraction c of the tumor sample has a normal copy-number (either by contamination of normal tissue or heterogeneity within the tumor), the copy-number ratio of this admixed sample will be $\rho' = c + \rho(1 - c)$. Thus, in heterogeneous samples, the only change to the method described above is the replacement of ρ by ρ'. The admixture rate c can be estimated from data by back-calculating c from empirical ρ' in LOH regions (see Appendix 2).

Segmentation and Sequential Merging

We used the circular binary segmentation (CBS) algorithm[30], as implemented in the R package DNAcopy [31], to subdivide the genome (exome). For each segment we combined the coverage by direct sum, and used mean coverage log ratio as the segment's log ratio log(R). Log ratio is used here to satisfy the input requirement of CBS algorithm. CNV call proceeds on each segment in the same manner as described above.

In order to achieve the most sensitive segmentation, we chose to start CBS with parameters that produce a large number of small segments, call CNV on the segments, and repeat the process with a smaller number of larger segments. We then merge the CNV segments sequentially, from finest segmentation to coarsest. By nature of CBS, finer segments are contained within coarser segments, and in merging step, we need to resolve conflicting CNV calls between finer segments within a larger segment. If a finer (smaller) segment has sufficient coverage to call a CNV event, the call persists. However, if it does not have sufficient coverage to reject the null hypothesis (i.e. not being called, or being called as copy-number neutral), a positive CNV call in a larger segment overrides the negative calls. An illustration of this sequential merging algorithm is given in the Appendix 2.

LOH Detection

First, we consider all polymorphic positions in the exome of the control sample, and for each of the positions, the B-allele count B_i is the number of reads with non-reference- or B-allele at that position. For a polymorphic position i, let N_i be the total number of reads mapped to that position (i.e. depth-of-coverage); thus the B-allele count B_i follows a binomial distribution

Binomial (p_i, N_i). A binomial that rejects the null hypothesis: $p_i = 0.5$ can be used to detect LOH at each polymorphic position.

Segmentation is done using CBS algorithm based on the absolute difference in B-allele frequencies $|BAF_{i,\text{case}} - BAF_{i,\text{control}}|$, where $BAF_i = B_i/N_i$. Within each segment, based on the realization that B-allele frequencies deviate from the null value of 0.5 under LOH, an F-test for equality of variance is used to detect significance increase in variance of BAF_{case} from that of BAF_{control} (other statistics were also considered, see Appendix 2). Finally, the LOH calls are merged sequentially as described above.

Exome Sequencing and Data Analysis

Exome Sequencing

High molecular weight whole genomic DNA from matched skin and tumor pair samples of a melanoma patient were sequenced on the Illumina Genome Analyzer (GAIIx). The UCLA Institutional Review Board (IRB) approved the collections of the DNA samples. The libraries were generated following the Agilent SureSelect Human All Exon Kit version 1.0.1 protocol, and the Illumina Genome Analyzer (GAIIx) flowcell was prepared according to the manufacturer's protocol. We performed three and four lanes of single-end sequencing for each of the skin and tumor samples, respectively, within the UCLA Center of High-throughput Biology (CHTB). The base-calling was performed by the real time analysis (RTA) software (version 1.6) provided by Illumina.

Sequence Data Analysis

Novoalign from Novocraft Short Read Alignment Package (http://www.novocraft.com/index.html) was used to align each lane's QSEQ file to the reference genome. Human Genome reference sequence (hg18, March 2006, build 36.1), downloaded from the UCSC genome database located at http://genome.ucsc.edu and mirrored locally, was indexed using novoindex program (–k 14 –s 3). The output format was set to SAM and default settings were used for all options. Using SAMtools (http://samtools.sourceforge.net/), the SAM files of each lane were converted to BAM files, sorted and merged for each sample and potential PCR duplicates were removed using Picard (http://picard.sourceforge.net/)[32]. To retrieve the depth of coverage information of each base, we generated a PILEUP file for each sample using SAMtools and calculated the average coverage per capture interval using a custom script. Here, we used processed BAM files that were used to call the single nucleotide variants (SNVs) while reducing the likelihood of using spuriously mismapped reads to call the variants: the last 5 bases were trimmed and only the reads lacking indels were retained [33]. The detailed description of the mutational landscape of this tumor sample is in preparation.

Genome-wide SNP Genotyping

Both the skin and the tumor samples were submitted to the Southern California Genotyping Consortium (SCGC) at UCLA for genotyping on the Illumina Omni-1 Quad BeadChip, which consists of 1,140,419 SNPs (1,016,423 genotyping probes and 123,996 CNV probes) distributed across the genome. The Illumina GenomeStudio V2010.1 Genotyping Module version 1.6.3 was used to calculate the B-allele frequency (BAF) values and the log R ratio (LRR) for each probe and the copy number aberration (CNA) and loss of heterozygosity

(LOH) of the autosomes were inferred from these values using the genoCN R package [34]. The genoCNA function was used with the default parameters.

Copy-Number Analysis using ERDS (Estimation by Read Depth with SNVs)

The same PILEUP file we used to generate the average coverage per capture interval was used to run ERDS [35]. The SNV file that is required by ERDS was generated by using SAMtools varFilter tool default parameters and SVA (Sequence Variant Analyzer) snp_filter.pl script. The result file generated by ERDS was summarized using the SVA software (http://people.genome.duke.edu/~dg48/sva/index.php).

Comparison between CNV Calling Methods

In assessing performance of ExomeCNV, we used all mapped exons as the sample space. CNV calls on other platforms were mapped to the exons and compared to calls by ExomeCNV. Thus specificity is the proportion of copy-neutral exons correctly identified by ExomeCNV, while sensitivity is the proportion of amplified (or deleted) exons correctly identified. Similarly, LOH performance is assessed using all polymorphic positions as the sample space.

Results

ExomeCNV for CNV and LOH Detection

ExomeCNV uses a normalized depth-of-coverage ratio approach to identify CNV and LOH from exome sequencing information of paired case/control samples (for example, paired tumor/normal) in a way that optimizes sensitivity and specificity. We begin by assuming that although there are potentially exon-specific biases due to laboratory capture methods and sequence-specific biases, these are independent of sample and so are nearly uniform for a

particular exon across samples. As a result, simply assessing the ratio of depth-of-coverage of each exon reduces such bias (see Appendix 2).

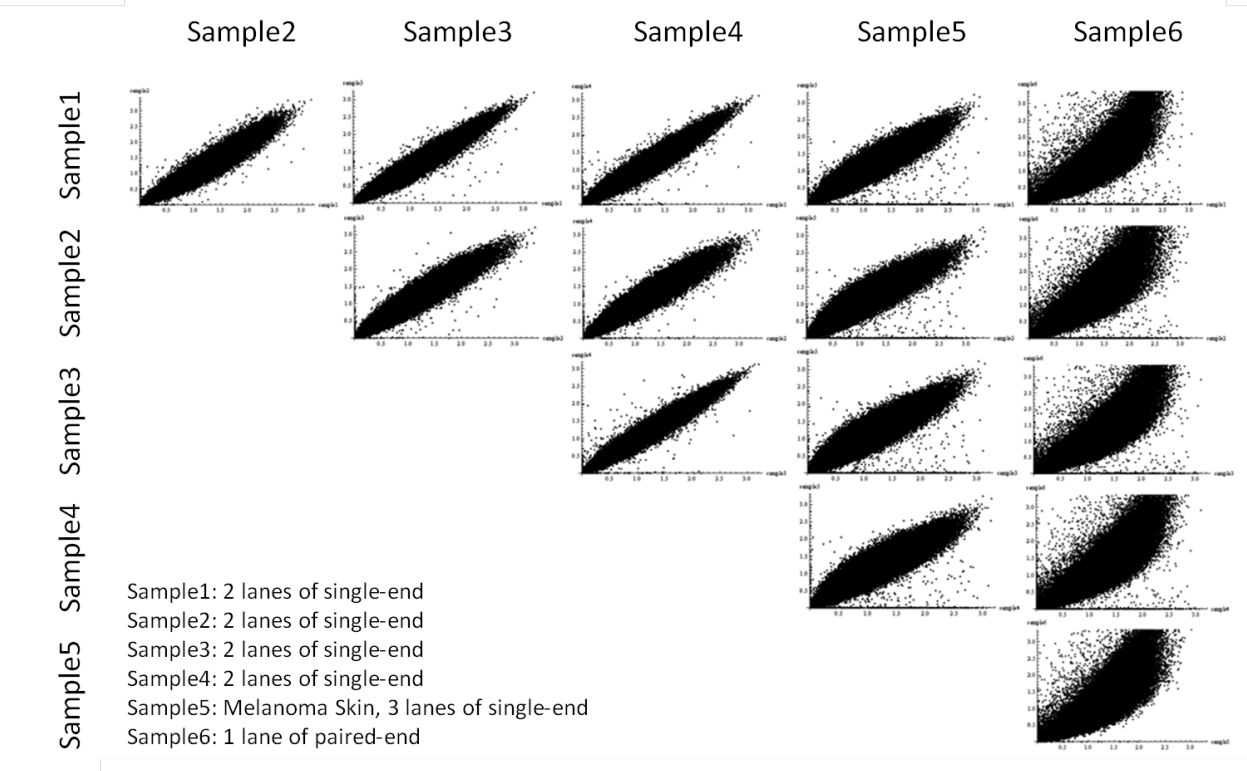


Figure 2.2 Correlation of Depth-of-Coverage across exome sequencing samples. To demonstrate the consistency of capture and sequencing efficiency of individual exons represented by the depth-of-coverage per exon, the normalized individual exon coverage in all pairs of 6 independent exomes were plotted. All six samples were captured using the Agilent SureSelect Human All Exon. Samples 1-5 had mean base coverages of 36~39X as a result of 2 (samples 1-4) or 3 (sample 5) lanes of GAIIx single-end sequencing per sample. Sample 6 had mean base coverage of 60X as a result of 1 lane of GAIIx paired-end sequencing and demonstrates substantially different biases in individual exonic depth-of-coverage.

To establish validity of this fundamental assumption, we compared depth-of-coverage of exons across five independent samples from five different subjects (Samples 1-5 in Figure 2.2). All samples were captured using the same probe set (Agilent SureSelect Human All Exon G3362) and sequenced at mean base coverages of 36-39X as a result of two (Samples 1-4) or

three (Sample 5) lanes of GAIIx single-end sequencing per sample (see Methods). As shown in Figure 2.2, a high correlation was observed among the five samples (Pearson correlation 0.908-0.975, mean=0.947, sd=0.027), arguing for the validity of our assumption.

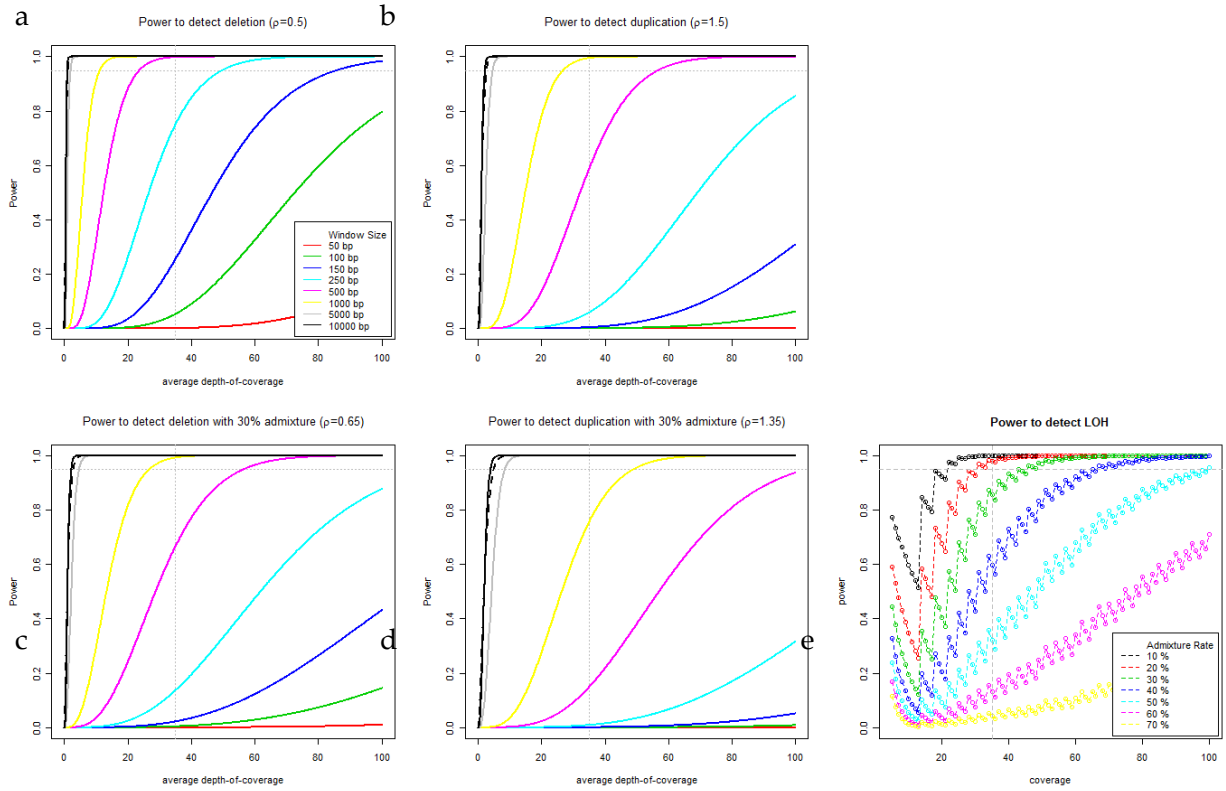


Figure 2.3 Examples of the power of ExomeCNV to detect segmental duplication, deletion, and LOH based on an analytical calculation. Power is plotted relative to mean depth-of-coverage in the genomic segment, setting false positive to 1 per genome based on an analytical model of genome-wide power of detection at different window sizes (inset, panel a-d). Windows are the total length of a given sequence at a given exon or the sum of length of exons adjacent to each other in the genome. The effect of admixture (rate of 30%) on the power to detect deletions and single copy duplications are shown in (c) and (d), respectively. Panel (e) plots the power of LOH detection versus depth-of-coverage of individual polymorphic position (single base pair) with variable admixture rates (inset). The periodicity of the power curve is due to discrete nature of the binomial test. The 35x depth of coverage is chosen because it is a typical minimal average depth of coverage for exome sequencing and is thus a conservative view of power within typical exome sequencing datasets.

The same level of consistency was not observed when single-end data were compared with paired-end data (Sample 6; Pearson correlation 0.855-0.877, mean=0.871, sd=0.009) due to the lack of independence between pairs of reads in paired-end data. Thus, care must be taken to ensure consistency of library preparation and sequencing method between samples used in analysis. Here, all of our analyses used exome sequencing data from a melanoma (Sample 5) and a matched normal skin, both processed and sequenced in the same manner (see Methods).

Analytic power calculation of exonic CNV and LOH detection

For each exon, the number of sequencing reads aligning within it appears to follow a Poisson with mean directly proportional to the size of the exon and the copy-number (see Appendix 2), but assuming that we have sufficiently deep coverage, we can approximate this by a normal distribution with mean equal to variance. We apply the Geary-Hinkley transformation [27-29], which converts a ratio of two normally distributed variables to a standard normal distribution, and a CNV is identified by a significant deviation of the transformed ratio from the null, standard normal distribution (see Methods). Allowing only one false positive per genome, we analytically determine the statistical power of this CNV detection approach for different depth-of-coverage, and the results are shown in **Figure 2.3**A-B. For detecting deletions, 95% power is achieved for segments of size 500 bp or more (**Figure 2.3**A), while detection of a single copy duplication is achieved with 95% power for segments of size 1,000 bp or more (**Figure 2.3**B) with a mean segmental base coverage of 35x. We note that the power of the method improves substantially with higher depth-of-coverage, and an individual exons deletion/duplication status would be more powerfully observed by including additional flanking intronic sequence in the capture probe design. Genomic DNA admixture, as

expected, diminishes power, but even with 35x coverage of a given exon, a length of greater than 1000bp is observed over 95% of the time. Exons or segments captured with 500bp at target sequence are observed at 95% power only with greater than 55x base coverage. A more thorough consideration of specificity and receiver operating characteristic (ROC) curves are produced in the Appendix 2.

To estimate LOH, we focused on the non-reference-allele or "B-allele" frequency (BAF) of polymorphic positions in the sequenced regions. The observed B-allele count at a polymorphic position can be modeled by a binomial distribution with depth-of-coverage as sample size and the probability of observing a B-allele proportional to the B-allele copy-number, which is equivalent to the LOH state. Because the expected value of BAF at a normal (non-LOH) polymorphic position is 0.5, a significant deviation of BAF from 0.5 identifies LOH. With sufficient depth-of-coverage, LOH can be detected at a single polymorphic position.

The effect of sample admixture

The specificity and sensitivity of this CNV detection method depends not only on the depth-of-coverage but also the rate of admixture, whereby non-mutated genomes contaminate mutated genomes in the sampled tissue/cells. In the absence of admixture, the average depth-of-coverage ratio is 0.5 for deletion, 1.5 for one-copy duplication, and the BAF at an LOH site is either 0 or 1; however, in cancer biopsy sequencing, this is rarely observed in practice due to admixture with normal or non-mutated tumor cells. Thus, we assess the effect of admixture ranging from 10-70%, which is frequently observed in tumor samples (**Figure 2.3**E). With admixture, the depth-of-coverage ratio and the BAF will tend to the null values of 1 and 0.5,

respectively, making the CNV and LOH detection harder. **Figure 2.3**C-D show a reduction of power in detecting deletion (c) and one-copy duplication (d) as a result of 30% admixture. There is an approximate two-fold increase in the size of the exonic sequence detectable in the presence of 30% non-mutated genomic DNA. Capping the false positive rate at 0.001 and assuming 35x mean depth-of-coverage, a power curve (**Figure 2.3**E) shows 0.95 sensitivity of detecting LOH at a single polymorphic position with admixture up to 30%.

Using circular binary segmentation to merge exonic CNV/LOH

Because CNV and LOH can, and usually do, span multiple exons, we extended our method above to call CNV/LOH on larger segments derived from summing data of sequentially spaced exons in the human genome. We apply circular binary segmentation (CBS) [30,31] to subdivide the genome and then combine depth-of-coverage of exons and BAF of polymorphic positions within each segment, composed of arbitrary number of individual exons, to search for larger CNV and LOH. In the case of CNV, since reads are independent of each other, the sum of depth-of-coverage of all exons in a segment constitutes the segment's depth-of-coverage, and the CNV test can be performed as described above. In the case of LOH, since B-alleles are not always on the same chromosome, BAF cannot be combined by direct summation. Instead, since BAF deviates from the null value 0.5 under LOH, a significance increase in variance of BAF from control (F-test for equality of variances) indicates LOH (several other statistics were also considered, see Appendix 2). Finally, we repeated the process of CBS and CNV/LOH-calling, ranging granularity of segmentation from finest to coarsest, and merged the CNV/LOH calls by prioritizing positive calls of finer segments over coarser ones (see Methods for details). In the case of our melanoma sample, we performed CBS/sequential merging at five levels of

granularity and observed 165,130 merging events in the first iteration followed by 121, 79, 105, and 66 in the subsequent iterations for a total of 165,501 merging events.

Validation

To test the performance of ExomeCNV we analyzed exome sequencing data from a melanoma and a matched normal skin (Appendix 2); the average depth-of-coverage of the data is 42.8x for the tumor and 37.5x for the normal sample, which are sufficient to achieve at least 90% sensitivity and specificity based on the power calculation above. We first estimated the false positive rate of the algorithm by calling CNVs on two sequencing lanes of the same normal tissue library, treating one as case and the other as control; any CNV call from this would be false positive. Our method correctly called most exons as non-CNV. In particular, setting p-value thresholds to ensure minimum specificity of 0.9, 0.99, and 0.999, we observed specificity of 0.916, 0.995, and 1.0, respectively (Appendix 2). Furthermore, we tested the sensitivity of ExomeCNV by analyzing copy-number of sex chromosomes in a pair of male and female exome data that were available internally (see Methods). Using the male exome as control, ExomeCNV correctly identified female chromosome X as being "duplicated" and chromosome Y as being "deleted" (Appendix 2) with no false negatives.

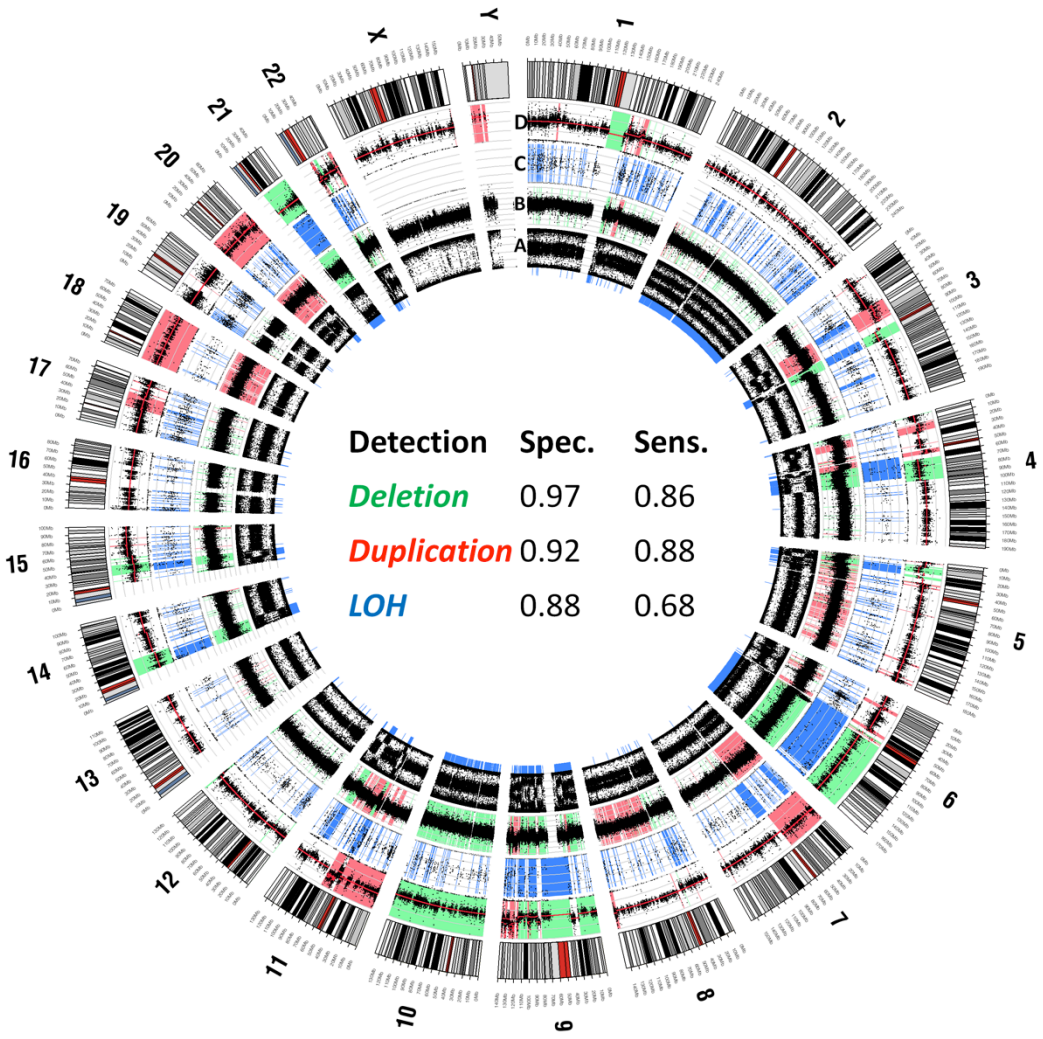


Figure 2.4 Analysis of Melanoma and Paired Normal Samples. Interpretation of deletion, duplication, and LOH from exonic sequence data using ExomeCNV and plotted with Circos. The most outer ring shows the chromosome ideograms in a pter-qter orientation, clockwise with the centromeres in red. From inside to outside, each data track represents (A) B-allele-frequency (BAF) from Omni-1 genotyping array with the region of LOH highlighted in blue underneath the track; (B) Log R Ratio (LRR) from genotyping array with the region of gain highlighted in red and the region of loss highlighted in green; (C) BAF from ExomeCNV output from ~40x depth-of-coverage exome sequencing with the region of LOH highlighted in blue; (D) log ratio of tumor and normal depth-of-coverage with the segment mean in red line, the region of gain highlighted in red and the region of loss highlighted in green. The LOH and CNV for the chromosome Y were not called for the genotyping data as genoCN (the algorithm used to call CNV from Omni-1) is not designed to analyze chromosome Y. The table in the middle summarizes best achievable specificity and sensitivity of ExomeCNV in detecting CNV and LOH relative to CNV/LOH calls from Omni-1 array assessment.

We then used ExomeCNV to predict CNV and LOH in the melanoma samples and compared our results to those obtained from Illumina Omni-1 Quad Beadchip genotyping array assessment of the same samples (**Figure 2.4** and Appendix 2). The sizes of the CNV segments from ExomeCNV range from single exon (120bp) to whole chromosome (chr 10 and 18) (size distribution of CNV calls is presented in the Appendix 2). Treating calls from the genotyping array experiment as a standard, ExomeCNV had 97% specificity and 86% sensitivity for detecting deletions, 92% specificity and 88% sensitivity for detecting amplifications, and 88% specificity and 68% sensitivity for detecting LOH even though there is substantial variability across the genome. Higher depth-of-coverage from the sequence data for each exome would likely further improve concordance. We note that this is a dramatic improvement over the ERDS [35] CNV caller which, when applied to these data, achieves 16% sensitivity and 83% specificity for deletion and 50% sensitivity and 56% specificity for amplification (see Circo plot showing results from the three methods in the Appendix 2). For CNV segments called by ExomeCNV but not by the genotyping arrays, we found that most lie within regions in which there is a low density of genotyping markers; thus the false positive rates (and the associated specificity) for ExomeCNV here may in fact be lower.

Discussion

The resolution of CNV detection with ExomeCNV is limited largely by the probe design. The CNV segments identified by our method range from 120bp (single exons with higher than average coverage) to 240Mb in size (whole chromosomes); however, the true breakpoint can be anywhere in the space between the terminal exon called within a CNV region and the adjacent exon in a non-CNV region. Hence, although a given CNV event can be detected at a single exon

in some instances, the absolute resolution of our method is in fact limited to the inter-exon distance around an exon, which can be as small as 125bp or as large as 22.8Mb with the median of 5kb (statistics based on SureSelect Human All Exon Kit G3362).

Although ExomeCNV relies on the availability of matched control samples, we can also derive a matched control sample from a pool of other samples which then serves as an effective control. This is useful for the identification of germline inherited or *de novo* CNVs in an individual. Because the expected copy number in the reference population is constant (usually two), by the law of large numbers, averaging depth-of-coverage from sufficiently many samples yields a good control set, assuming that they are all captured using the same probe set and capture method and sequenced in the same manner. This may limit the application of ExomeCNV to data generated at a given site with a given protocol. Calling CNVs using this pooled sample as background will generate CNV calls that are present in the case sample but not the control population. Also, by the central limit theorem, pooling independent samples helps reduce variance in depth-of-coverage and increases precision of our method. We have pooled as few as 8 samples and have observed that this is indeed the case (Appendix 2). However, it is important to note that using pooled sample as control imposes a strong assumption that the samples do not share common CNV regions and that the population has an average genomic copy number of two. Other potential challenges of using the pooled sample as control are discussed in the Appendix 2.

Because ExomeCNV depends on an estimate of the admixture rate c, misspecification of c would affect its performance. We performed sensitivity analysis and found that

misestimating *c* would have a strong effect on sensitivity and specificity of CNV detection. Fortunately, LOH detection provides some data to directly estimate *c*, as LOH detection does not depend on a prior knowledge of *c* (Appendix 2). For the melanoma sample, our estimate of 30% admixture rate matches that from genotyping arrays, confirming the validity of this approach. However, there are advantages to slightly overestimating *c* as it makes the method more conservative and reduces false positives.

As we have shown, CNV and LOH detection is readily possible from exome sequencing data, extending the utility of this powerful approach. The fundamental basis that makes this approach possible is the consistency of depth-of-coverage of each exon (and BAF by extension) across multiple samples for each individual exon, as demonstrated in 5 samples performed in our lab (Appendix 2, **Figure 2.2**). This consistency permits reliable parametric modeling of the shift in depth-of-coverage and BAF distributions, hence accurate identification of CNV and LOH. However, we do not observe the same level of consistency when comparing depth-of-coverage across different library types. For instance, a sixth sample was performed using a paired-end approach which results in very different coverage of each exon (**Figure 2.2**), and as a result, ExomeCNV does not perform well when the control sample library is of one type and the case is of another, or when the case and control have significantly different coverage levels. Resolving these issues is a work-in-progress.

From the analytical power calculations, assuming 35x coverage (which is the lower end of a reasonable amount of sequence for variant calling and easy to generate with a variety of technologies), CNV detection has a limit of about 500bp (in transcript coordinates), which is

typically equivalent to 2-3 exons and spans about 10kb of genomic space on average. Increased depth-of-coverage, which is likely to become the norm as sequencing costs decrease, reduces the interval size that is reliably detectable and should push the method to single exonic deletion resolution. Currently, CNV and LOH information should be detectable in whole-exome sequencing data at a resolution that is almost equivalent to what one can obtain from a dense SNP genotyping array.

ExomeCNV is available as a CRAN package "ExomeCNV".

Chapter 2 Bibliography

1. Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet.

2. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 106: 19096-19101.

3. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.

4. Chou LS, Liu CS, Boese B, Zhang X, Mao R (2010) DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. Clin Chem 56: 62-72.

5. Hoischen A, Gilissen C, Arts P, Wieskamp N, van der Vliet W, et al. (2010) Massively parallel sequencing of ataxia genes after array-based enrichment. Hum Mutat 31: 494-499.

6. Raca G, Jackson C, Warman B, Bair T, Schimmenti LA (2010) Next generation sequencing in research and diagnostics of ocular birth defects. Mol Genet Metab 100: 184-192.

7. Volpi L, Roversi G, Colombo EA, Leijsten N, Concolino D, et al. (2010) Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. Am J Hum Genet 86: 72-76.

8. Rehman AU, Morell RJ, Belyantseva IA, Khan SY, Boger ET, et al. (2010) Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. Am J Hum Genet 86: 378-388.

9. Johnston JJ, Teer JK, Cherukuri PF, Hansen NF, Loftus SK, et al. (2010) Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. Am J Hum Genet 86: 743-748.

10. Choi CH, Lee KM, Choi JJ, Kim TJ, Kim WY, et al. (2007) Hypermethylation and loss of heterozygosity of tumor suppressor genes on chromosome 3p in cervical cancer. Cancer Lett 255: 26-33.

11. Deng FY, Zhao LJ, Pei YF, Sha BY, Liu XG, et al. (2010) Genome-wide copy number variation association study suggested VPS13B gene for osteoporosis in Caucasians. Osteoporos Int 21: 579-587.

12. Fong CT, Dracopoli NC, White PS, Merrill PT, Griffith RC, et al. (1989) Loss of heterozygosity for the short arm of chromosome 1 in human neuroblastomas: correlation with N-myc amplification. Proc Natl Acad Sci U S A 86: 3753-3757.

13. Jankowska AM, Szpurka H, Tiu RV, Makishima H, Afable M, et al. (2009) Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. Blood 113: 6403-6410.

14. Sha BY, Yang TL, Zhao LJ, Chen XD, Guo Y, et al. (2009) Genome-wide association study suggested copy number variation may be associated with body mass index in the Chinese population. J Hum Genet 54: 199-202.

15. Shuin T, Kondo K, Torigoe S, Kishida T, Kubota Y, et al. (1994) Frequent somatic mutations and loss of heterozygosity of the von Hippel-Lindau tumor suppressor gene in primary human renal cell carcinomas. Cancer Res 54: 2852-2855.

16. Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61: 437-455.

17. Wain LV, Pedroso I, Landers JE, Breen G, Shaw CE, et al. (2009) The role of copy number variation in susceptibility to amyotrophic lateral sclerosis: genome-wide association study and comparison with published loci. PLoS One 4: e8175.

18. Kato M, Kawaguchi T, Ishikawa S, Umeda T, Nakamichi R, et al. (2010) Population-genetic nature of copy number variations in the human genome. Hum Mol Genet 19: 761-773.

19. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, et al. (2010) Diversity of human copy number variation and multicopy genes. Science 330: 641-646.

20. Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, et al. (2010) Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet 42: 385-391.

21. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464: 704-712.

22. McCarroll SA (2010) Copy number variation and human genome maps. Nat Genet 42: 365-366.

23. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207-211.

24. Urban AE, Korbel JO, Selzer R, Richmond T, Hacker A, et al. (2006) High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proc Natl Acad Sci U S A 103: 4534-4539.

25. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, et al. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods 6: 99-103.

26. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19: 1586-1592.

27. Geary RC (1944) Extension of a theorem by Harald Cramer on the frequency distribution of the quotient of two variables. Journal of the Royal Statistical Society 107: 56-57.

28. Geary RC (1930) The frequency distribution of the quotient of two normal variates. Journal of the Royal Statistical Society 93: 442-446.

29. Hinkley DV (1969) On Ratio of 2 Correlated Normal Random Variables. Biometrika 56: 635-&.

30. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557-572.

31. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657-663.

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

33. Clark MJ, Homer N, O'Connor BD, Chen Z, Eskin A, et al. (2009) U87MG Decoded: The Genomic Sequence of a Cytogenetically Aberrant Human Cancer Cell Line. PloS Genetics 6.

34. Sun W, Wright FA, Tang Z, Nordgard SH, Van Loo P, et al. (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Res 37: 5365-5377.

35. Zhu M, Need, A., Ge, D., Singh, A., Feng, S., Maia, J., Cirulli, E., Heinzen E., Fellay J., Ottman R., Milner J., Shianna, K. and Goldstein, D. (2010) Detection of copy number variation using whole genome sequence data from twenty human genomes. In Preparation.

Chapter 3: Microbial Co-Occurrence Relationships in the Human Microbiome

Abstract

The healthy microbiota show remarkable variability within and among individuals. In addition to external exposures, ecological relationships (both oppositional and symbiotic) between microbial inhabitants are important contributors to this variation. It is thus of interest to assess what relationships might exist among microbes and determine their underlying reasons. The initial Human Microbiome Project (HMP) cohort, comprising 239 individuals and 18 different microbial habitats, provides an unprecedented resource to detect, catalog, and analyze such relationships.

Here, we applied an ensemble method based on multiple similarity measures in combination with generalized boosted linear models (GBLMs) to taxonomic marker (16S rRNA gene) profiles of this cohort, resulting in a global network of 3,005 significant co-occurrence and co-exclusion relationships between 197 clades occurring throughout the human microbiome. This network revealed strong niche specialization, with most microbial associations occurring within body sites and a number of accompanying inter-body site relationships. Microbial communities within the oropharynx grouped into three distinct habitats, which themselves showed no direct influence on the composition of the gut microbiota. Conversely, niches such as the vagina demonstrated little to no decomposition into region-specific interactions. Diverse mechanisms underlay individual interactions, with some such as the co-exclusion of Porphyromonaceae family members and *Streptococcus* in the subgingival plaque supported by

known biochemical dependencies. These differences varied among broad phylogenetic groups as well, with the Bacilli and Fusobacteria, for example, both enriched for exclusion of taxa from other clades. Comparing phylogenetic versus functional similarities among bacteria, we show that dominant commensal taxa (such as Prevotellaceae and *Bacteroides* in the gut) often compete, while potential pathogens (e.g. *Treponema* and *Prevotella* in the dental plaque) are more likely to co-occur in complementary niches. This approach thus serves to open new opportunities for future targeted mechanistic studies of the microbial ecology of the human microbiome.

Introduction

In nature, organisms rarely live in isolation, but instead coexist in complex ecologies with various symbiotic relationships [1]. As defined in macroecology, observed relationships between organisms span a wide range including win-win (mutualism), win-zero (commensalism), win-lose (parasitism, predation), zero-lose (amensalism), and lose-lose (competition) situations [2-4]. These interactions are also widespread in microbial communities, where microbes can exchange or compete for nutrients, signaling molecules, or immune evasion mechanisms [4-6]. While such ecological interactions have been recently studied in environmental microbial communities [7-10], it is not yet clear what the range of normal interactions among human-associated microbes might be, nor how their occurrence throughout a microbial population may influence host health or disease [11].

Several previous studies have identified individual microbial interactions that are essential for community stability in the healthy commensal microbiota [12-15], and many are further implicated in dysbioses and overgrowth of pathogens linked to disease [16]. Each

human body site represents a unique microbial landscape or niche [17,18], and relationships analogous to macroecological "checkerboard patterns" [3] of organismal co-occurrence have been observed due to competition and cooperation [19-22]. For example, dental biofilm development is known to involve complex bacterial interactions with specific colonization patterns [23-25]. Likewise, disruption of relationships among the normal intestinal microbiota by overgrowth of competitive pathogenic species can lead to diseases, e.g. colonization of *Clostridium difficile* in the gut [26]. However, no complete catalog of normally occurring interactions in the human microbiome exists, and characterizing these co-occurrence and co-exclusion patterns across body sites would elucidate both their contributions to health and the basic biology of their ecological relationships. Thus, characterizing key microbial interactions of any ecological type within the human body would serve as an important first step for studying and understanding transitions among various healthy microbial states or into disease-linked imbalances.

As has been also been pointed out in macroecology, however, the analytical methodology needed to comprehensively detect such co-occurrence relationships is surprisingly complex [27]. Most existing studies employ simple measures such as Pearson's or Spearman's correlation to identify significant abundance relationships [13,15,28]. These methods are suboptimal when applied without modification to organismal relative abundances [29]. Since absolute microbial counts are not known and measurements depend on sampling and sequencing depth, an increase in one relative abundance must be accompanied by a compositional decrease in another, leading to spurious correlations among non-independent

measurements [30]. In addition, sparse sequence counts can cause artefactual associations for low-abundance organisms with very few non-zero observations [29]. Conversely, association methods such as log-ratio based distances [31] that have been developed specifically for such compositional data are difficult to assign statistical significance, a vital consideration in high-dimensional microbial communities containing hundreds or thousands of taxa.

Here, we have addressed these issues to catalog a baseline of normal microbial interactions in the healthy human microbiome. The Human Microbiome Project (HMP) [32] sampled a disease-free adult population of 239 individuals, including 18 body habitats in five areas (oral, nasal, skin, gut, and urogenital), providing 5,026 microbial community compositions assessed using 16S rRNA gene taxonomic marker sequencing [32]. We have developed a suite of methods to characterize microbial co-occurrence and co-exclusion patterns throughout the healthy human microbiome while suppressing spurious correlations. Specifically, these were 1) an ensemble approach including multiple similarity and dissimilarity measures, and 2) a compendium of generalized boosted linear models (GBLMs) describing predictive relationships, both assessed nonparametrically for statistical significance while mitigating the effects of compositionality. Together, these methods provide a microbiome-wide network of associations both among individual microbes and between entire microbial clades.

Among the 726 taxa and 884 clades in the HMP data, we examined both intra-body site and inter-body site relationships as a single integrated microbial co-occurrence network. Each relationship represents co-occurrence/co-exclusion pattern between a pair of microbes within or between body sites among all subjects in the HMP (in contrast to studies within single subjects

of microbial co-occurrences across biogeography, e.g. [33,34]). This ecological network proved to contain few highly connected (hub) organisms and was, like most biological networks, scale-free. Co-occurrence patterns of the human microbiome were for the most part highly localized, with most relationships occurring within a body site or area, and there were proportionally few strong correspondences spanning even closely related body sites. Each pair of organisms was assessed for positive (e.g. cooperative) or negative (e.g. competitive) associations, and in many cases these patterns could be explained by comparing the organisms' phylogenetic versus functional similarities. In particular, taxa with close evolutionary relationships tended to positively associate at a few proximal body sites, while distantly related taxa with functional similarities tended to compete. The resulting network of microbial associations thus provides a starting point for further investigations of the ecological mechanisms underlying the establishment and maintenance of human microbiome structure.

		Houston			St. Louis		
Body Area/Site	**Total**	**Total**	**Female**	**Male**	**Total**	**Female**	**Male**
Oral	*3022*	*2038*	*840*	*1198*	*984*	*456*	*528*
Buccal mucosa	340	228	92	136	112	53	59
Hard palate	334	221	90	131	113	53	60
Keratinized gingival	337	226	95	131	111	51	60
Palatine Tonsils	340	225	92	133	115	54	61
Saliva	309	227	94	133	82	35	47
Subgingival plaque	341	228	92	136	113	53	60
Supragingival plaque	349	232	97	135	117	55	62
Throat	321	219	92	127	102	46	56
Tongue dorsum	351	232	96	136	119	56	63
Gut	*351*	*228*	*94*	*134*	*123*	*58*	*65*
Stool	351	228	94	134	123	58	65
Airways	*282*	*190*	*82*	*108*	*92*	*37*	*55*
Anterior nares	282	190	82	108	92	37	55
Skin	*921*	*554*	*233*	*321*	*367*	*159*	*208*
Left Antecubital fossa	158	85	37	48	73	25	48
Right Antecubital fossa	160	83	33	50	77	34	43
Left Retroauricular crease	303	198	87	111	105	50	55
Right Retroauricular crease	300	188	76	112	112	50	62
Urogenital	*450*	*286*	*286*	*0*	*164*	*164*	*0*
Mid vagina	149	93	93	0	56	56	0
Posterior fornix	150	95	95	0	55	55	0
Vaginal introitus	151	98	98	0	53	53	0
Total	***5026***	***3296***	***2230***	***2324***	***1730***	***1292***	***1184***

Table 3.1: 16S rRNA gene sequencing data from the Human Microbiome Project used to assess microbial co-occurrence relationships in the human microbiome. We considered microbial associations in a total of 5,026 samples from the Human Microbiome Project (HMP) comprising 18 body sites in 239 individuals recruited at two clinical centers (Baylor College of Medicine, Houston, TX and Washington University at St. Louis, MO), which in total contained 726 reliably detectable bacterial phylotypes. For details of HMP samples and data processing, see [32].

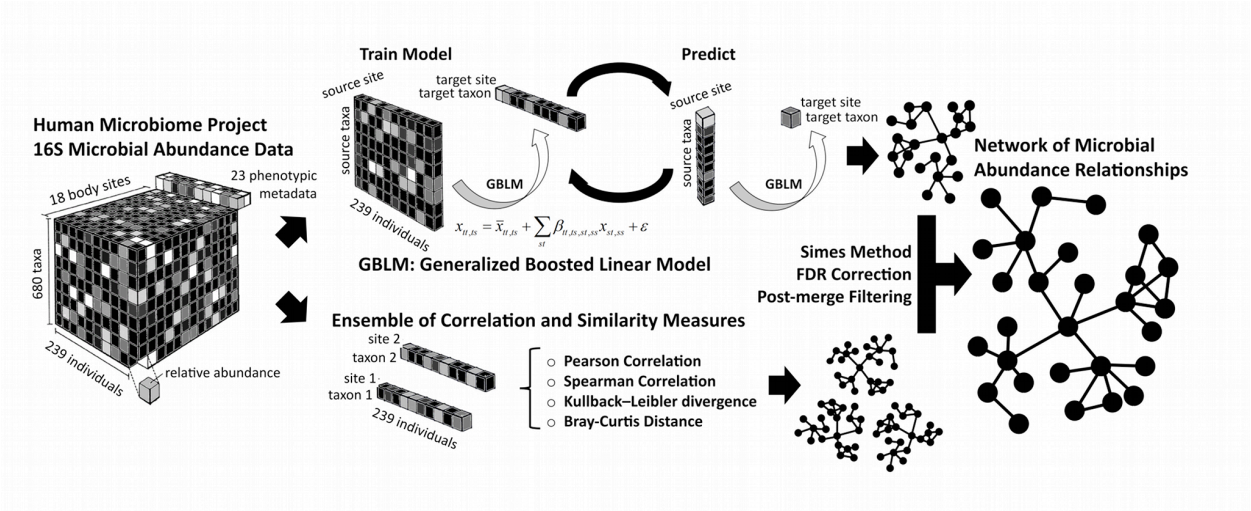


Figure 3.1: Methodology for characterizing microbial interactions using a compendium of similarity measures. 16S data from the Human Microbiome Project (HMP) were collected from 18 body sites in a cohort of 239 healthy subjects and assessed using 16S rRNA gene sequencing. We analyzed microbial co-occurrence and co-exclusion patterns in these data by developing two complementary approaches: a compendium of Generalized Boosted Linear Model (GBLMs) and an ensemble of similarity and dissimilarity measures. Each approach produced a network in which each node represented a microbial taxon within one body site, and each edge represented a significant association between microbial or whole clade abundances within or across body sites. The resulting association networks produced by each individual method were merged as p-values using Simes method, after which FDR correction was performed. Associations with FDR q-values >0.05, inconclusive directionality, or fewer than two supporting pieces of evidence were removed. This provided a single global microbial association network for taxa throughout the healthy commensal microbiota.

Methods

Two complementary approaches, an ensemble of multiple similarity/dissimilarity measures and a compendium of generalized boosted linear models (GBLMs), were used to interrogate significant associations between microbial abundances. These were drawn from 18 body sites assayed by the Human Microbiome Project at two clinical centers using 16S rRNA gene sequencing. Simes method and Benjamini-Hochberg-Yekutieli false discovery rate (FDR) correction were used to combine the resulting networks. From this merged, global network, we summarized overall network properties (degree distribution, modularity, etc.), assessed patterns of microbial connectivity within and among body sites, and identified highly connected (hub) microbes. Phylogenetic and functional distances were calculated based on 16S rRNA gene sequence similarity and shared orthologous gene families, respectively, and combined with the network.

Please see Appendix 3 for an extensive discussion of the methodology used to assess relationship significance in compositional data, which is presented in summary below.

16S data acquisition and processing

The 16S rRNA gene-based dataset of the normal (healthy) human microbiome was made available through the Human Microbiome Project (HMP) and is detailed in [32]. Briefly, it consists of 454 FLX Titanium sequences spanning the V1 to V3 and V3 to V5 variable regions obtained for 239 healthy subjects enrolled at clinical sites in Houston, TX and St. Louis, MO. These cover 18 body sites covering five areas: the oral cavity (nine sites: saliva, tongue dorsum, palatine tonsils, keratinized gingiva, hard palate, buccal mucosa, throat, and sub- and

supragingival plaques), the gut (one site: stool), the vagina (three sites: introitus, mid-vagina, and posterior fornix), the nasal cavity (one sample: anterior nares), and the skin (four sites: left and right antecubital fossae and retroauricular creases). Sequences of both 16S windows were processed separately using mothur [35] into phylotypes using the RDP taxonomy as described in [32] and [36], with full protocols also available on the HMP DACC website (http://hmpdacc.org/HMMCP). Genus level and above phylotypes were used for this analysis, for which the datasets from both windows were combined.

This resulted in more than 5,000 samples comprising 910 taxa made available as part of the HMP (http://hmpdacc.org/HMMCP). These were further processed for this study by excluding any phylotype not supported by at least two sequences in at least two samples. Samples were removed as suspect if the most abundant taxon was detected by fewer than 1% of the sequences supporting it in the sample in which it was most abundant, and counts for the remaining 726 taxa were converted to relative abundances in each of the resulting 5,026 samples. Due to potential differences between clinical centers, the dataset was conservatively split into two subsets for further analysis, subjects recruited in Houston (3,296 samples) and those recruited in St. Louis (1,730 samples).

Generalized Boosted Linear Models

GBLM definition and construction

For each resulting dataset, a compendium of generalized boosted linear models (GBLMs) was constructed by selecting all 324 combinations of source body sites *ss* and target sites *ts*. Each GBLM was fit using the abundances of all source taxa *st* within the source site to

predict the abundance of each target taxon *tt* within the target site using a sparse linear model of the form:

$$x_{tt,ts} = \bar{x}_{tt,ts} + \sum_{st} \beta_{tt,ts,st,ss} x_{st,ss}$$

All additional non-leaf clades in the RDP [37] taxonomy (i.e. families, orders, etc. up to the bacterial and archaeal domains) were included as source and target taxa. For *ss* equal to *ts*, i.e. predicting the abundance of a taxon *tt* when the abundances of all taxa in the same body site are known, the abundances of all parent and descendant clades of *tt* were excluded from the available source taxa *st*. That is, when *ss*=*ts* and *tt* was of the form *domain*|*phylum*|*class*|...|*clade*, all source taxa of the form *domain*, *domain*|*phylum*, *domain*|*phylum*|*class*, etc. or of the form *domain*|*phylum*|*class*|...|*clade*|*subclade*, *domain*|*phylum*|*class*|...|*clade*|*subclade*|*subsubclade*, etc. were excluded from the source taxa *st*. This prevented the abundances of $x_{tt,ts}$ from being predicted using abundances $x_{st,ss}$ on which they were directly dependent, while allowing the detection of predictive relationships between distinct clades within the same body site.

The linear model was generalized to include binary categorical target taxa (in this case only gender and ethnicity) using standard logistic regression:

$$\text{logit}(x_{tt,ts}) = \bar{x}_{tt,ts} + \sum_{st} \beta_{tt,ts,st,ss} x_{st,ss}$$

As this is clearly an extremely high-dimensional problem, multiple a priori and post hoc steps were taken to enforce model sparsity and to avoid overfitting for each (*ss*, *ts*, *tt*) tuple. The first of these was to exclude from the available *st* any taxon not correlating with *tt* at a

Spearman correlation of nominally p<0.05. The second was to boost linear model fitting rather than attempt to fit all $\beta_{tt,ts,st,ss}$ simultaneously [38]. Boosted linear models retain the usual L2 least squares penalty, but are constructed in a manner similar to sparse forward variable selection or the LASSO [39]. βs are considered for inclusion in the model one at a time and the parameter minimizing sum of squared error selected and included. However, each subsequent round of parameter fitting operates on the residuals of all previous rounds, thus "upweighting" poorly fit examples, and the inclusion of further non-zero βs stops after a fixed number of iterations.

This tuning parameter and the model fitting process was 10-fold cross-validated and selected from the most accurate (by root mean square error for continuous *tt* and AUC for binary) of 50, 100, or 150 boosting iterations using the caret [40] and mboost [41] R packages. This resulted in a compendium of 7±4.9 non-zero parameters β, each evaluated with a 10x cross-validated R^2/AUC and a nominal R^2/AUC on the full dataset. Final model quality scores were assigned by A) subtracting AUCs below 0.5 from one, since caret does not calculate AUCs directionally, and B) retaining the minimum of the cross-validated and nominal R^2/AUC. Any continuous model not achieving an R^2 above zero after adjustment for the number of non-zero parameters ($AR^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$), parameters p, training samples n) was discarded (55,424 retained).

GBLM filtering and significance

Even from cross-validated goodness-of-fit scores, the compositional structure of relative abundance data prevents straightforward assessment of model significance (see Appendix 3). We thus additionally fit twenty models per (*ss*, *ts*, *tt*) tuple after bootstrap re-sampling the

values of *tt* across samples. The AR^2/AUC of these bootstrap models provided a confidence interval around the observed AR^2/AUC. Any model for which the 90% confidence interval failed to include the observed AR^2/AUC was discarded. To construct the null distribution of associations due to compositionality alone, we fit twenty additional models after permuting the values of *tt* across samples and renormalizing them sample-wise, thus retaining compositional effects but breaking true associations. The GBLM was re-fit and the resulting null distribution of AR^2/AUC values used to assess significance of the true model. The mean and standard deviation of the bootstrap distribution was z-tested against this null distribution using the jointly pooled standard deviation, providing one p-value per model per clinical center. Any model with FDR adjusted p-value greater than 0.05 was discarded (18,286 retained).

Ensemble scoring

Data preprocessing

The 16S data described above were first normalized by dividing each sample by its total phylotype sum. Mislabeled samples were removed [42] and samples were again processed as two subsets, one per clinical center (Houston and St. Louis). These were encoded as a matrix in which each row represented a phylotype in a specific body site and each column represented an individual during one sampling visit. Rows with more than 2/3 zero counts were removed, leaving matrices of 1,217 (Houston) and 1,408 (St. Louis) phylotype-bodysite composite features collected for 248 and 144 subject-visit points, respectively.

Ensemble score calculation

We built on composite co-occurrence scores as described in, for example, [43] (for protein functions) and [20] (graph clustering) to find groups of microbial lineages co-existing across a large number of environments. Specifically, we combined four diverse measures in order to overcome two major challenges in the inference of co-occurrence networks, particularly appropriateness of scoring measures to sparse count data and determination of statistical significance. The first is exemplified by the double-zero problem, in which a zero indicates either that an organism is absent or that it is below detection limit [27]; Pearson and Spearman correlations are sensitive to this, while the Bray Curtis dissimilarity is not. The latter issue arises from the need to normalize across samples with unknown absolute abundances, either by relativizing or by downsampling; either procedure results in constrained sample sums, which introduce artificial correlations [29].

We thus employed an ensemble approach combining four diverse measures: two measures of correlation (Pearson and Spearman) and two measures of dissimilarity (Bray-Curtis (BC) and Kullback-Leibler (KLD)). For BC and KLD calculations, rows were divided by their sum prior to computation. Additional measures were considered for our ensemble, including the Hellinger, Euclidean, variance of log-ratios and other measures, but these proved to be well-represented by the smaller final ensemble (see Supplemental Figure 3.6-3.7).

Ensemble network building

After running each of the above measures on the two 16S data matrices, one per clinical center, we set measure-specific thresholds as a pre-filter such that each measure contributed

1,000 top-ranking and 1,000 bottom-ranking edges to the network. Edge scores were computed only between clade pairs without parent-descendant relationship (e.g. without pairs of the type Actinobacteridae|Actinomycetales or Actinomycetales|Propionibacterineae) for clades in the same body site. To assign statistical significance to the resulting differently-scaled scores, we first computed edge- and measure-specific permutation and bootstrap score distributions with 1,000 iterations each. In order to address the compositionality issues discussed above [29], we re-normalized the data in each permutation, providing a null distribution that captures the similarity introduced by compositionality alone (see Appendix 3). We then computed the p-value as above by z-scoring the permuted null and bootstrap confidence interval using pooled variance. P-values were tail-adjusted so that low p-values correspond to co-presence and high p-values to exclusion. For BC and KLD, we did not compute re-normalized permutations, because these measures are intrinsically robust to compositionality [30]. Instead, we calculated their p-values using the bootstrap interval compared to a point null value that was computed by permutation.

Finally, to remove unstable edges, we removed all edges whose score was not within the 95% confidence interval (limited by the 2.5 and 97.5 percentiles) of the bootstrap distribution. Additionally, a small number of Bray-Curtis scores were manually curated and removed after generating erroneous negative relationships when applied to abundance profiles including only one extreme outlier. This affected only the clades for St. Louis: Actinomycetales in stool, *Corynebacterium* and Corynebacteriaceae in the tonsils, *Lactobacillus* and Lactobacillaceae in the anterior nares and for Houston: an unclassified *Neisseria* in the left retroauricular crease.

Mitigating the compositional effect in relative abundance analysis

Data summarized as relative abundances are referred to as compositional [30]. Because they sum to one, their elements are not independent and may exhibit spurious correlations regardless of the true underlying relationship. To mitigate this effect, we assessed the significance of our ensemble scores and GBLMs as described above, each by comparing a bootstrap confidence interval around the observed score with a permuted null distribution that includes repeated renormalization to account for compositional effects alone. In each permutation, only the target taxon row was randomized and all samples subsequently renormalized to a constant sum. Because permutation breaks correlation structure while renormalization reintroduces compositionality, a null distribution coupling these elements induces correlation from compositional structure alone. Comparing this null distribution to a standard bootstrap confidence interval around the observed value provided a straightforward nonparametric test of association accounting for compositionality. Simulation studies showed that the bootstrap-renormalization scheme was successful in discounting compositional effects while preserving true correlations (see Appendix 3).

Network merging by Simes method and FDR correction

Simes method was used to combine all ten networks (5 methods x 2 study centers) into one final network, as it is robust against non-independent tests [44]. A strict intersection of the two clinical centers' networks, rather than a p-value combination, was also examined and found to be over-stringent due to systematic differences in the data (Supplemental Figure 3.8). After merging, p-values on each final edge were corrected to FDR q-values using the Benjamini-Hochberg-Yekutieli method and a q-value cutoff of 0.05 was applied. The positivity or

negativity of each relationship was determined by consensus voting over all integrated data sources, ranging from -10 (most negative) to 10 (most positive, see Figure 3.2). Edges with indeterminate directionality (direction score of zero) were removed. Finally, only edges with at least two (out of ten) supporting pieces of evidence were retained.

Computation of network modularity

The formula by Clauset et al. [45] compares the fraction of edges within input clusters with the fraction of within-cluster edges that would be expected for a randomized network. We clustered the network using the Markov cluster algorithm (MCL) [46] and computed network modularities for a range of inflation parameters. The strongest modularity (0.28) was measured for an inflation of 1.3 and is slightly below the cut-off recommended by Clauset et al.. This modularity was, however, higher than any measured for 100 randomized networks (which preserved node and edge number, but in which edges were randomly re-assigned) and was therefore retained as significant.

Assessment of significant connectivity density within and among body sites and classes

To assess whether specific body sites were more connected than expected by chance, we repeatedly (1,000 times) selected as many nodes as a body site contains from the global network at random and counted their edge number. This resulted in a distribution of edge numbers for random node sets. To retain only plausibly significant edges for further calculation, we corrected for multiple testing by multiplied the nominal p-value from this distribution with the number of tests carried out and retaining values below 0.05. We repeated this test separately for positive, negative, intra- and cross-edges. For visualization, the network itself was also

separated into within-site, within-area, and between-area subsets for further inspection (Supplemental Figure 3.9).

Assessment of relationships between body-site-specific and class-specific clade groups

To assess the interaction strength between body site and clade groups (Figures 3.4 and 3.5), the number of relationships between nodes in each group pair was counted and normalized by the group member product, which represents the number of links two groups can potentially form. This was repeated in 1,000 randomized networks (generated as described for the computation of network modularity). A p-value was computed from the count distribution and node group relationships with p-values above 0.05 were discarded, retaining only the fractions of total possible relationships which were significantly higher than that expected by chance.

Phylogenetic and functional similarity scores

Phylogenetic distances

Genome sequences of 1,107 organisms were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, December 2010) and 16S sequences were extracted. These 16S sequences were aligned using MUSCLE 3.8.31 [47] and a full phylogenetic tree reconstructed by maximum likelihood using FastTree 2.1 [48]. A matrix of all pair-wise distances was created from this cladogram and distances between any two nodes (e.g. families, orders, etc.) calculated by taking the median of all distances (as provided in units from FastTree) between all pairs of leaf taxa descending from the two nodes.

Functional distances

Functional complements of the same genomes were summarized using COG [49] families as assigned by NCBI annotations. This resulted in an abundance matrix with 4,685 columns (corresponding to COG families) and one row per genome. Columns summing to less than 10% of the number of genomes were removed, resulting in 3,514 usable COG families. Pairwise scores between genomes were calculated using Jaccard index, with distances for higher-level clades computed using medians in the same manner as described above. Final functional distances were represented as the Jaccard index of non-shared COG families between pairs of genomes.

Results

We inferred a microbiome-wide microbial interaction network by analyzing 5,026 samples from the Human Microbiome Project (HMP) comprising 18 body sites, 239 individuals recruited at two clinical centers, and 726 bacterial phylotypes detected by 16S rRNA gene sequencing (Table 1). Our study aimed to determine co-occurrence and co-exclusion relationships among the relative abundances of microbial taxa across all individuals, potentially indicative of their ecological relationships. We thus combined two complementary approaches, namely an ensemble of multiple similarity and dissimilarity measures (henceforth "ensemble approach") and a compendium of generalized boosted linear models (GBLMs, henceforth "GBLM approach"). Both methods were applied to the HMP data to produce microbial interaction networks in which each node represented a microbial clade (taxon or group of taxa) connected by edges that were weighted by the significance of their association (positive or negative). Spurious correlations due to compositional structure of relative abundance data [29]

were prevented by a novel bootstrap and re-normalization approach assessing the degree of association present beyond that expected by compositionality alone. We used Simes method followed by Benjamini-Hochberg-Yekutieli false discovery rate (FDR) correction to combine the resulting networks (Figure 3.1). A detailed final network is provided in Supplemental Figure 3.1, with a comparison of all networks in Supplemental Figure 3.7 and additional information in Methods. This provided a single global microbial interaction network capturing 3,005 associations among 197 phylotypes, spanning all available body sites from the human microbiome (Figure 3.2).

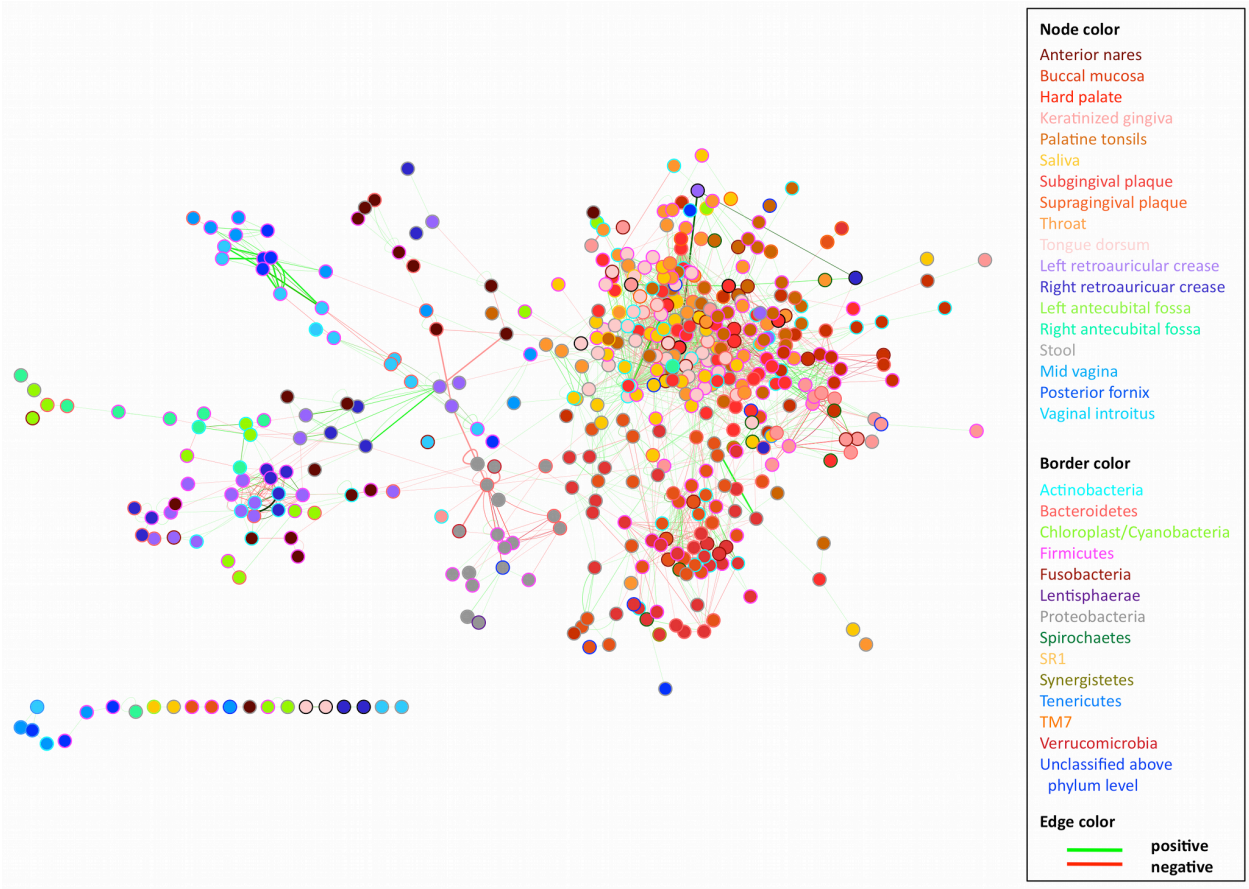


Figure 3.2: Significant co-occurrence and co-exclusion relationships among the abundances of clades in the human microbiome. A global microbial interaction network capturing 1,949 associations among 452 clades at or above the order level in the human microbiome, reduced for visualization from the complete network in Supplemental Figure 3.1. Each node represents a bacterial order, summarizing one or more genus-level phylotypes and family-level taxonomic groups. These are colored by body site, and each edge represents a significant co-occurrence/co-exclusion relationship. Edge width is proportional to the significance of supporting evidence, and color indicates the sign of the association (red negative, green positive). Self-loops indicate associations among phylotypes within an order; for a full network of all phylotypes and clades, see Supplemental Figure 3.1. A high degree of modularity is apparent within body areas (skin, urogenital tract, oral cavity, gut, and airways) and within individual body sites, with most communities forming distinct niches across which few microbial associations occur.

A global network of microbial co-occurrence and mutual exclusion within and among body site niches of the human microbiome

Global properties of the microbiome-wide network of microbial associations are summarized in Figures 3.2 and 3.3. A dominant characteristic of the network was its habitat-specific modularity. After grouping the 18 body sites into five broad areas (oral, skin, nasal, urogenital, and gut), the large majority of edges were found clustered within body areas (98.54%), and these clusters were sparsely connected through a minority of edges (1.46%). This is confirmed by the network's high modularity coefficient of 0.28 (as defined by [45]) and Markov clustering of the network (see Methods and Supplemental Figure 3.2). It has long been observed that sites within the human microbiome are distinct in terms of microbial composition [50], and this proved to be true of microbial interactions as well: microbial relationships within each body area's community were largely unique (Table 2). The microstructure of interaction patterns - and thus in the underlying ecology - was different for different areas, however. For example, all vaginal sites within the urogenital area were interrelated in a single homogeneous community, whereas interactions within the oral cavity suggested microbial cross-talk among three distinct habitats [51]. This can be observed quantitatively based on the proportions of microbial interactions spanning body sites within each area, e.g. 69.57% among the vaginal sites and 53.19% among the oral sites, both exceeding the microbiome-wide baseline. The skin was further unique in that the large amount (57.65%) of its associations related microbes in corresponding left and right body sites (left and right antecubital fossae and retroauricular creases), reflecting consistent maintenance of bilateral symmetry in the skin microbiome.

Edges	# of edges	Percent
Within same body area	2961	98.54%
Within same body site	1409	(47.59%)
Among skin sites	196	6.52%
Between left and right skin sites	113	(57.65%)
Within the airways (anterior nares)	31	1.03%
Among oral sites	2598	86.46%
Between different oral sites	1382	(53.19%)
Within the gut	67	2.23%
Among vaginal sites	69	2.30%
Between different vaginal sites	48	(69.57%)
Total	3005	

Table 3.2: Summary statistics of microbial associations in the normal human microbiota. Microbial co-occurrence and co-exclusion relationships summarized within the five major body areas and relationships spanning different body sites within these areas. Percentages are fractions of the total number of edges in the network, while percentages in parentheses represent fractions of edges within each body area.

We began decomposing the network by categorizing microbial associations within each body area into body-site-specific relationships of two types: cross-site and within-site interactions. On average, these two classes make up 53.11 and 46.89 percent of the total edges, respectively (Table 3.2). First focusing on cross-site associations, a majority (66.10%) of such relationships were co-occurrences between the same or taxonomically related clades in proximal or bilateral body sites. This reflects coordinated community structure among ecologically related niches, such as similar dental plaques, vaginal sites, and bilateral skin sites. Body sites specifically connected by many positive associations were either in direct contact (e.g. tongue and saliva), proximal (e.g. sub- and supragingival plaques), or similar in terms of environmental exposure (e.g. bilateral skin sites), thus providing mechanisms to support comparable microbiota and exhibiting high levels of microbial co-occurrence. This pattern held true for the minority (33.90%) co-exclusions as well, with many occurring between bilateral skin

sites or within subgroups of the oral cavity ([51], see also Figure 3.5). This suggested that the first level of hierarchical co-occurrence structure in this network corresponded with groups of body sites representing distinct microbial habitats.

Conversely, within-site relationships showed a much more balanced ratio of microbial co-occurrence (48.26%) vs co-exclusion (51.74%) interactions. Many of the negative within-site relationships were associated with the abundant signature organisms characteristic of each body site [52], for example *Streptococcus* in the oral cavity and *Bacteroides* in the gut. The relative abundances of these signature taxa varied greatly among individuals, in some cases (e.g. *Bacteroides*) spanning from 1% to 97% within a body site across the HMP population. It is generally very difficult to determine from relative abundance measurements alone whether these negative associations represent true anti-correlation (e.g. one organism out-competing another) or overgrowth of one organism while the rest of the population remains unchanged (resulting in a negative correlation due to compositionality of these data). This problem has a long history in quantitative ecology [29,30,53]. Our methods generally determine these relationships in the human microbiome to be stronger than what would be expected from compositionality alone (see Methods and Supplemental Methods), and the negative interactions detected here are thus likely biologically informative. This is supported by the fact that they are strongest in cases where distinct alternative dominant community members occurred among different individuals (e.g. Prevotellaceae vs. Lactobacillaceae in the vaginal area [54] or *Propionibacterium* vs. *Staphylococcus* on the skin [52,55]). The increase in negative interactions within habitats is also in line with the fact that most competitive mechanisms require proximity

or physical contact [56], whereas positive interactions are likely to also occur from microbiome-wide shared environmental exposures.

Association properties globally and within body sites demonstrate the basic ecological organization of the human microbiota

We further assessed several other measures of network community structure. Globally speaking, the network followed a scale-free degree distribution typical of biological systems, meaning that most clades possessed few interactions but a few clades possessed many (Figure 3.3A [57]), The network had a low average path length of three (contrasted with six in randomized networks), meaning that short paths existed between most clades [58], and it possessed a low average per-node cluster coefficient (0.1) measuring the local density of connections. Together, these values indicate that the microbial association network is structured to be scale-free and thus robust to random disruption [57], with only sparse local multi-organism clusters. Since these data only describe phylotypes at approximately the genus level, it remains to be seen whether a greater degree of locally clustered functional associations emerges among Operational Taxonomic Units (OTUs), species, or strains within these phylotypes. As the cluster coefficient distribution was not well described by the inverse node degree distribution [59], the network possesses no strong hierarchical modularity despite its scale-freeness, in contrast to the strong habitat-centric modularity.

The diversity of microbial interactors (i.e. number of unique phylotypes) within each body site also proved to directly dictate its interaction density (Figure 3.3B). That is, communities with a greater number of different organisms had a proportionally greater number

of positive and negative associations. Within these sites, the number of relationships scaled directly with the number of unique phylotypes (adjusted R^2 of 0.75), the only body site with more interactions than expected for its diversity being the tongue. This site also harbored the top-ranking hub phylotype (Firmicutes, see Figure 3.3A). In combination with the behavior of specific microbial hubs as discussed below, this might argue that most microbial taxa form strong metabolic or functional associations with adjacent taxa inhabiting the same body site habitat, allowing consortia to specialize within highly localized microbial niches [50].

When randomizing between rather than within body sites, no body site pairs possessed more cross-site associations than expected (with the slight exception of tongue dorsum), whereas most body sites were significantly enriched for within-site relationships (the only exceptions being posterior fornix, mid-vagina, and antecubital fossae, which tended toward too few phylotypes to reach significance; see Figure 3.3D), again confirming the microbiome's habitat-driven modularity. When calculating network properties in a body-area-specific manner, we found that the overall average path length between nodes in the oral cavity, which contributes most of the samples, was much larger (~3.4) than those of the other body areas (ranging from ~1.1 to ~2.0). In addition to supporting the aforementioned degree of inter-site habitat formation in the oral cavity, this intriguingly suggests that other body sites in which fewer samples are currently available (see Table 3.1) have not yet exhausted the detection of microbial relationships in the human microbiome. More samples and greater sequencing depth may further improve detection power.

Key taxa including members of the Firmicutes act as network hubs coordinating many relationships throughout the microbiome

We next examined the associations of individual clades with respect to interaction degree, observing highly connected "hub" clades to be found within each body area. Two classes of hubs appeared in the association network: clades highly connected within one body site, and clades acting as "connectors" between multiple body sites. Hubs included both specific taxa (e.g. *Porphyromonas*, see Figure 3.3A) and larger taxonomic groupings (e.g. the phylum Firmicutes). Within-site hubs were often, although not always, abundant signature taxa (detailed below), high-degree exceptions including *Atopobium* on the tongue (28 total associations, 16 within-site) and *Selenomonas* on both tooth plaques (20 total/19 within and 7 total/3 within for supra- and subgingival, respectively). The latter provides a striking example of the niche-specificity of these low-abundance within-site interactors, as *Selenomonas* averages only 1.1% and 1.2% of the sub- and supragingival plaque communities, respectively, but associates preferentially (20 of 27, 74%) with members of the greater oxygen availability supragingival community. The clade's detection as a within-site hub thus corresponds with the ecology that might be expected of an organism known to be oxygen-sensitive, fastidious, and grown best in co-culture [60].

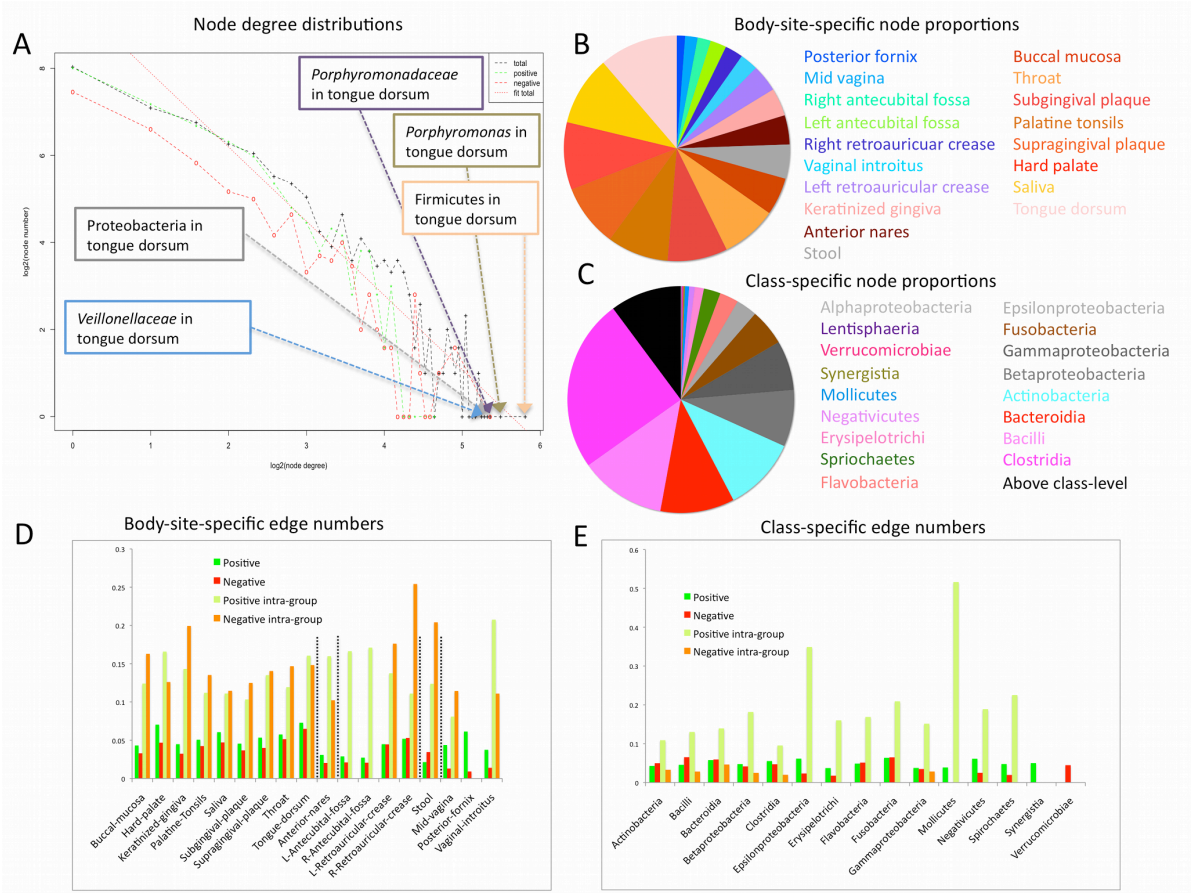


Figure 3.3: Global network properties summarizing key microbial hubs and interaction patterns. A) Node degree distributions of overall, co-occurrence, and co-exclusion associations in the human microbiome. This is well-fit by a power law with slope -1,7 (dotted red regression line, adjusted R^2=0.9). Node degree indicates the number of links that connect a node to others in the network. Power law degree distributions, referred to as scale-free, mean that most nodes have only a few edges and are often connected by a few high-degree hub nodes, The top five most connected hubs as indicated in callouts, mainly signature oral taxa including *Porphyromonas* in the tongue dorsum. B) and C) Node proportions after division of the network into body sites (B) or classes (C). Both pie charts show that the composition of the network (in agreement with underlying data) is skewed towards the oral cavity (B) and its constituent Firmicutes (including Bacilli and Clostridia) (C). (B) further agrees with published measures of body sites' alpha diversity [61]. D) and E) Composition of relationships among microbes grouped according to body site (D) and taxonomic class (E). In E), the first two bars (green and red) include the fraction of all possible edges incident to at least one node representing a class or one of its members (root scaled for visualization). The second two bars (lime and orange) only include pairs of microbes that are members of the same class, again normalized as a fraction of total possible interactions and root scaled. The Bacilli, Bacteroidia, and Fusobacteria contain significantly more negatively associated microbes than expected by permutation testing, and classes overall are depleted for negative associations, indicating that members of the same class tend not to compete strongly with each other in these communities.

Between-site hubs typically operated among body sites within the same area as described above, with two of the five most connected hub clades in the network falling into this connector category linking multiple body sites, Firmicutes and Proteobacteria on the tongue (see Figure 3.3A). The Firmicutes and *Porphyromonas* (phylum Bacteroidetes) hubs in the tongue also had the largest numbers of negative connections among all phylotypes, and all of these highly interactive clades centered on the tongue and spanned multiple related oral habitats. Signature clades such as the Firmicutes are of course highly functionally diverse, and this network suggests that the few abundant members in any one habitat [52] might instead serve as "information processors" throughout a body area. In contrast to the low-abundance within-site hubs, this would allow them to provide baseline functionality complemented by distinct, less abundant clades with which they co-occur within differing body site habitats.

Correspondingly, Firmicutes and other inter-site hub nodes showed a higher connectivity than the clades with highest intra-site degree (e.g. Bacteroidales in the subgingival plaque). Such clades with unusually frequent inter-site associations are thus outliers relative to the network's overall habitat-specific trend and suggest that inter-site hubs are particularly critical for associating similar sites within the same body area. In the oropharynx, for example, *Streptococcus* spp. with a modest degree of functional variation might be present throughout the habitat, interacting with distinct, more specialized clades within each body site [13]. Almost all such high-connectivity hubs occurred among oral sites (e.g. *Porphyromonas*, *Streptococcus*, *Veillonella*, and others), the first notable exception being the *Propionibacterium* hub on skin sites (left and right retroauricular crease). All of these follow the same pattern, however, in which

abundant phylotypes likely possessing within-clade functional diversity are distributed among related habitats within each individual.

Marked differences in ecological behavior between phylogenetic clades

We additionally examined the phylogenetic rather than biogeographical distribution of these associations, testing whether clades tended to support more phylogenetically related (within-clade) or diverse (between-clade) interactions. We first investigated purely quantitative degree distributions by summarizing clades at the class level. Associations were summarized as the fraction of all possible interactions that were observed to occur, separated into positive and negative bins (Figure 3.3E). In addition, clade-specific over-representation of these bins was tested for significance by randomization (see Methods). The only classes that showed significantly more negative (and, simultaneously, cross-clade) associations than expected were the Bacteroidia, Bacilli, and Fusobacteria. Most of the common classes in the human microbiome had more intra-clade edges than expected by chance (Actinobacteria, Bacilli, Bacteroidia, Betaproteobacteria, Clostridia, Epsilonproteobacteria, Fusobacteria, Gammaproteobacteria, Mollicutes, and Spirochaetes), most of which also have high cluster coefficients (Supplemental Figure 3.3). Taken together with the biogeographical interactions assessed above, the enrichment for within-class associations likely indicates a phylogenetic aspect of the same behavior. Specifically, if one member of such a class is abundant in one body site within an individual, it (or closely related class members) also tends to be enriched in related body sites.

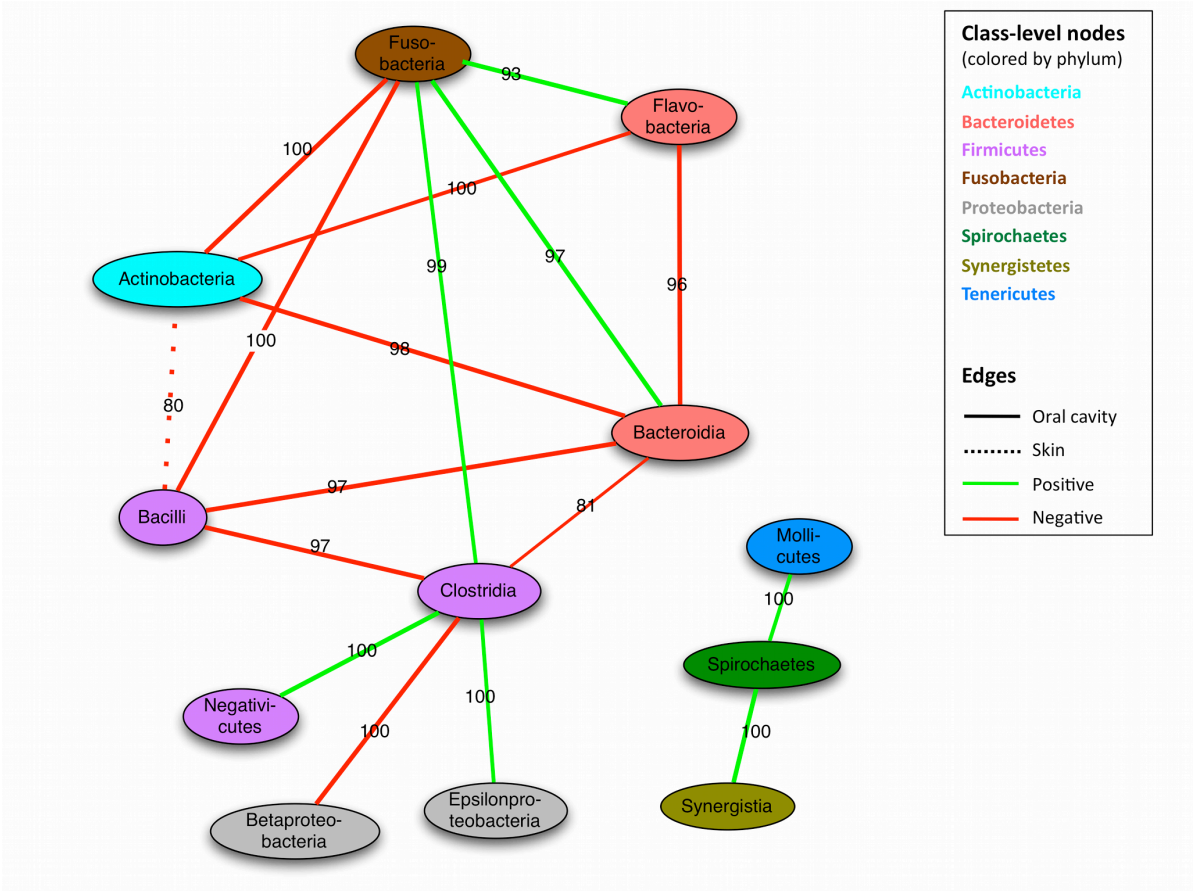


Figure 3.4: Co-occurrence of microbial clades within and among body areas. Nodes represent microbial classes colored by phylum, with edges summarizing aspects of their interactions over all body sites. Classes are linked when the number of edges between them is significantly larger than expected (randomization $p<0.05$, see Methods). Edge type (solid or dashed) indicates the body area contributing the most edges to the total interactions between two classes, with the label specifying the percentage contributed by this dominant body area. For instance, 80% of the edges between Bacilli and Actinobacteria come from skin sites. Green indicates co-occurrence, red exclusion. Most inter-class interactions occur in the mouth, with the Actinobacteria and Bacilli forming negative hubs.

We next considered relationships between class-level clades throughout the microbiome, summarized in Figure 3.4. Surprisingly, the Actinobacteria and Bacilli form only co-exclusion relationships with other classes, most strongly with Bacteroidia and Fusobacteria, and primarily within the oral cavity. These clades (which include the extremely abundant streptococci) might thus be largely self-sufficient in the functional diversity needed to maintain an oral community, excluding other clades when appropriately supported by e.g. environmental factors. Although a

few classes were linked by positive as well as negative interactions (e.g. Clostridia and Bacteroidia), none of these reached significance on randomization, although such behavior might suggest either that the clades exhibit co-occurrence only in some environments or that some members of the two classes co-occur while others co-exclude. As the oral communities are both the most data-rich and the most alpha-diverse in the human microbiome [52], it is not surprising that most relationships are observed within and among them. For instance, 97% of the specific mutual exclusions between Bacilli and Bacteroidia members occur in oral sites, as do 81% of the members of the Clostridia and Bacteroidia. The second largest contribution to the latter exclusion (~18%) comes from the gut, reflecting the frequently discussed Bacteroides/Firmicutes ratio observed in Western populations [15,62], and similar tradeoffs (with few positive associations) were observed in other habitats such as the skin (e.g. *Staphylococcus* in the Bacilli and *Propionibacterium* in the Actinobacteria [55]).

Co-exclusions such as these have previously been observed in the human microbiota to induce distinct alternative community configurations, which may differ across persons [15,54] as well as time points (e.g. early and late colonizers in community establishment or repopulation after disturbance). Although our methodology does not explicitly describe alternative community configurations, co-occurrence networks can in some cases capture them as extreme exclusion relationships between key microbial taxa. For instance, Ravel et. al reported five different vaginal communities in an independent cohort of healthy women, four dominated by Lactobacilli and the fifth diverse and featuring members of the Actinobacteria, Clostridia, Bacteroidia, and other classes. These alternative configurations occur as mutual

exclusions in our genus-level phylotypes between *Lactobacillus* and members of this fifth diverse community (particularly anaerobes such as Anaerococcus and the Prevotellaceae). Furthermore, we see a strong negative correlation in stool samples between *Bacteroides* and members of the gut community, including the Ruminococcaceae and other Firmicutes. In other body sites, the clade relationship network (Figure 3.4) features a negative interaction between Bacilli and Bacteroidia classes that mostly occurs in the oral cavity, and oral Porphyromonas (a member of the Bacteroidia) is among the most highly connected negative hubs. *Porphyromonas* is abundant (avg. 3.3% s.d. 3.9%) in oral habitats but not in most cases the dominant clade; the clade also includes potential oral pathogens [63], and this may be one of the more striking examples of functional competition and co-exclusion occurring with a specific clade among several oral communities.

Microbial relationships within digestive tract niches including *Fusobacterium* and *Prevotella* support known microbiology

The digestive tract is home to one of the most diverse and densely populated microbial communities in the human body [11]. Oral sites made up half of the body sites surveyed here, as well as exhibiting the greatest within-subject microbial diversity [52]. Correspondingly, associations between microbes within and among oral sites likewise comprised the majority (86.46%) of all edges in our co-occurrence network, also forming its largest connected component. This consisted of two clusters of organisms from the mouth soft tissues (gingiva, mucosa, and palate) and distal areas (tongue, throat, tonsils, and saliva); the oral hard surfaces (sub- and supra-gingival plaques) formed an additional isolated habitat that showed significantly fewer microbial associations with the remainder of the oral cavity (Figure 3.5). A

complementary analysis of the HMP microbiomes has revealed evidence of three sub-habitats within the oral cavity based on overall similarity of their microbial communities [51], and these results demonstrate that the shared community structures of these habitats were to a lesser degree recapitulated in terms of specific microbial associations (see Figure 3.5 below).

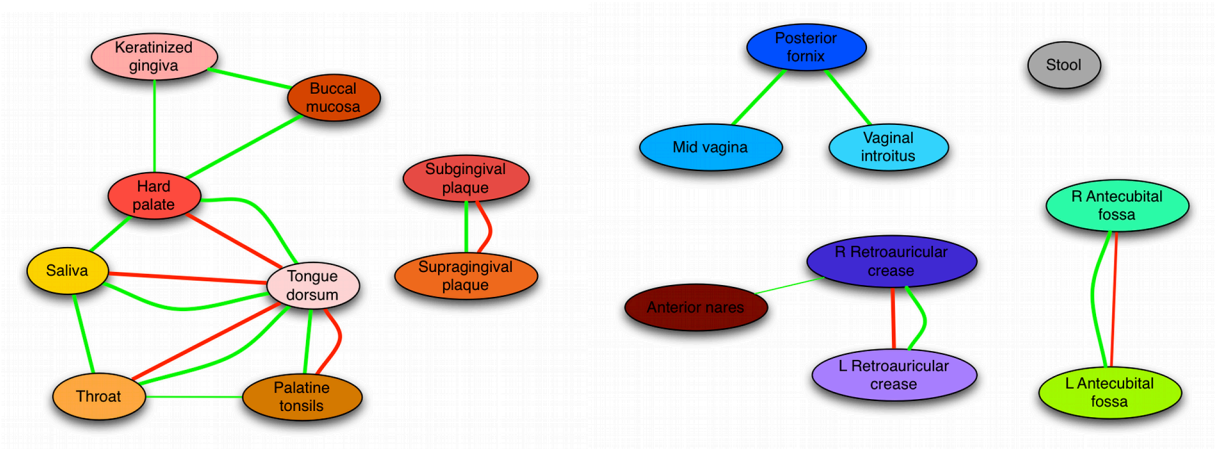


Figure 3.5: Related microbial niches as determined by associations spanning habitats at multiple human body sites. Each node represents a body site, with edge width indicating significant cross-site correlations (randomization $p<0.05$, see Methods). Green edges show co-occurrence, red co-exclusion. Skin, vaginal, oral soft tissue, and tooth plaque moieties are apparent, with the gut and airways notably lacking significant interactions with other available body site niches. However, most relationships between microbial relative abundances occur specifically within, rather than between, individual body sites.

Although the current study is associative and does not by itself establish causative mechanisms of interaction for these microbial associations, many that we detect in the oral cavity in particular are supportd by known metabolic or biochemical interactions. For instance, in the context of cell to cell interaction, *Fusobacterium* species are known to be bridging organisms in the development of oral biofilms by co-aggregation through physical contact [64]. This bridging occurs during biofilm maturation, allowing a more complex use of resources including sugars (the predominant carbon source for early colonizers) and proteins (used by

late colonizers). In the hard palate, for example, positive associations were found between *Fusobacterium* and *Capnocytophaga*, *Peptosptreptococcus*, and *Porphyromonas*, which are in agreement with previously published cell-to-cell interactions [65,66], and these predictions additionally implicate *Leptotrichia* and *Parvimonas*. Dental plaque associations included *Parvimonas*, *Prevotella*, and *Treponema*, also in agreement with existing evidence [67]. However, those previously published aggregations are strain specific and, this study may be observing broader effects than the direct cell-cell contact preferences in previously described associations.

Conversely, metabolic shifts may explain negative associations detected between other co-habiting microbes, e.g. *Tannerella* and *Streptococcus* in the subgingival plaque. The anaerobic *Tannerella* requires a much lower pO_2 than *Streptococcus* and is proteolytic, while *Streptococcus* is a saccharolytic colonizer of the tooth surface that uses sugars as its primary source of carbon and is oxygen tolerant [68,69]. This continuous nutritional, metabolite (e.g. hydrogen peroxide), and oxygen gradient between the supragingval and the subgingival biofilms, along with differential exposure to host factors in saliva, is reflected through the gradual drop of the abundance of *Tannerella* as the streptococci increase (Supplemental Figure 3.4). A similar example can be found in the *Prevotella* and species from the Flavobacteriaceae (represented here by *Capnocytophaga*; mean abundance 1.68±2.76%) in the tonsils. Less exposed surfaces of tonsillar crypts offer an anaerobic micro-environement favoring species like *Prevotella*, while other areas support the growth of carbon dioxide-dependant *Capnocytophaga*, a tradeoff that we detect here as a specific negative association.

Phyla such as the TM7 and Synergistetes have only recently been characterized at the genetic level in the oral cavity [70,71], and little is yet known about their roles in this microbial ecosystem. We identified a number of novel co-occurrences between members of these under-characterized phyla, including a positive association between members of the TM7 phylum (mean abundance 0.62±1.14%) and *Moryella* genus members (mean abundance 0.29±0.47%) in the tongue dorsum and a positive relationship between members of the Synergistetes phylum and *Treponema* in the subgingival biofilm. Since limited data on metabolic byproducts or requirements for these clades in the oral community are available, these newly identified putative interactors provide specific hypothesis for follow-up studies (e.g. by co-culture experiments).

The degree to which microbial shedding from the oral cavity along the digestive tract might seed the distal commensal gut microbiota is as yet unclear [72]. We found few (7) relationships between organisms in the two areas meeting our significance criteria, none of which were consistently supported by a majority of available data (Supplemental Figure 3.5), suggesting no such direct microbial seeding within our level of detection in the healthy adult microbiome. Interactions detected within the gut itself consisted primarily of negative associations between *Bacteriodes* and *Clostridia*, especially members of the Ruminococcaceae family. These negative relationships reflect the tradeoff between *Bacteroides* (mean abundance 48.79±22.94%, range 1.47-97.14%) and Firmicutes (mean abundance 27.04±16.52%, range 1.49-91.78%), the two dominant gastrointestinal taxa and the subject of previous close study [15,73]. While oral microbial transit is clearly important during founding of the microbiome in infancy

and in extreme cases such as illness [74,75], these data suggest that it occurs at low levels in the normal adult microbiome. In such hosts, the naturally dense microflora of the lower gut may serve to further exclude the few bacteria that survive gastrointestinal transit [72].

Similarities among niches in the microbiome determined by microbial associations spanning body sites

It is common practice to group microbial communities by ecological similarity [50,52], and we extended this analysis method by summarizing relationships among similar habitats based on microbial cross-talk (Figure 3.5). Specifically, we organized pairs of body sites by the frequency with which they demonstrated co-occurring (or excluding) microbes (see Methods). Overall, this network recapitulates similarities in community structure among these microbial habitats as assessed by beta-diversity [52], with the added information of which microbes might drive these similarities. Conversely, co-exclusions spanning multiple habitats might represent cases in which competitive relationships or differing responses to host environment might bridge multiple habitats. Stool microbes (representing the gut microbiota), as above, did not demonstrate any detectable associations with inhabitants of the mouth; the airways microbiota (nares) likewise associated minimally with other body site, although they were detectably structurally similar to the skin communities. The sub- and supra-gingival plaques were distinct from other mouth sites, and the vaginal communities and skin were again all highly similar. The sparsity of this body site network again illustrates that phylotypes rarely participate in detectable ecological relationships spanning distal body site habitats.

Functional and phylogenetic similarities among associated organisms suggest competitive and adaptive explanations for interactions

We hypothesized based on previous findings in environmental communities [20] that patterns of microbial co-occurrence and exclusion might be explained by their evolutionary relatedness and functional similarity. For example, closely related microbes might compete for limited resources, while functionally complementary bacteria would exhibit mutualism. To test this hypothesis, we compared two genomic properties of all microbial clades appearing in our network, their phylogenetic similarity (i.e. evolutionary relatedness) and a "functional" similarity score based on counting shared orthologous gene families (i.e. a measure of shared pathways and metabolic capacity). Phylogenetic distances were calculated as evolutionary divergence based directly on 16S sequence dissimilarity between all pairs of microbes. We compared this with a "functional" distance calculated as the Jaccard index of non-shared COG families between all pairs of microbial genomes (see Methods). For most pairs of microbes, these measures were highly correlated (Figure 3.6), not necessarily surprising in that both are influenced by gradual sequence change driven by molecular evolution.

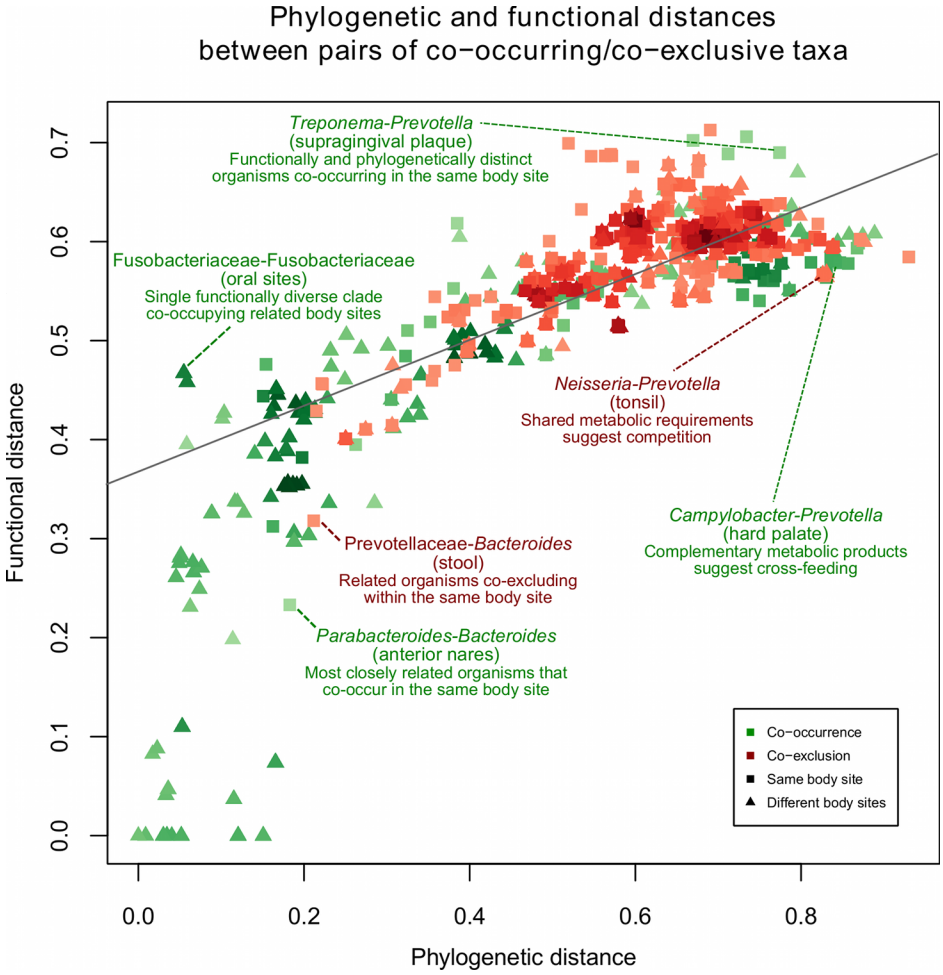


Figure 3.6: Functional and phylogenetic similarities between co-occurring microbes. Evolutionary (phylogenetic) distances among microbial clades were compared to the clades' functional potentials as defined by the Jaccard index of orthologous gene (COG) families shared between genomes (see Methods). Each point represents a pair of significantly associated microbes colored by direction of the association (green positive, red negative) and shaped by the type of relationship (triangle: between body sites, square: within site). Phylogenetic distances were inferred by FastTree [48] using species-level 16S sequences. Most interactions lie along the diagonal, reflecting the baseline correlation between these functional and evolutionary distances, with highly related clades co-occurring among related habitats (e.g. bilateral skin sites, proximal oral sites) in the lower left. Off-diagonal examples include potential competition among dominant gut signature taxa (e.g. Prevotellaceae/*Bacteroides*) and functional complementarity between distinct oral pathogens (e.g. *Treponema*/*Prevotella*).

However, several exceptions to this pattern were apparent among the interacting organisms of our study. First, a dramatic separation of phylogenetic and functional distances occurred between positively and negatively associated clades (Figure 3.6, green lower left vs. red upper right): positive associations were enriched for both phylogenetic and functional similarity, while negative associations showed the inverse pattern. This was partially explained by the basic observation that similar organisms occupy similar niches, as most relationships among similar organisms occurred between clades at different body sites and often between the same clade at two proximal (e.g. oral) or bilateral sites (e.g. left and right retroauricular creases). Conversely, the preference for negative correlations to occur between phylogenetically and functionally different organisms (top right) suggests that the wide range of co-exclusion mechanisms, not only direct competition but also toxin production, environmental modification, and differential niche adaptation [76] required substantial time to develop throughout evolution. Furthermore, interactions in the same body site were primarily negative, suggesting that competition or subniche differentiation were more prevalent in these data than were collaboration or niche sharing.

Exceptions to both of these trends did occur, however, in that related organisms occasionally showed within-site competition, and phylogenetically distant clades sometimes co-occurred. A highlighted example of the former was the negative association between *Bacteroides* and Prevotellaceae family members (also phylum Bacteroidetes) in the gut, reflecting the recurrent tradeoff of this genus with the *Prevotella* as previously linked to enterotypes [15] and/or dietary patterns [77]. As these organisms are closely related, this might reflect alternative

metabolic specializations in an otherwise fairly similar gut environment. Conversely, the *Aggregatibacter* were positively associated with members of the highly dissimilar Flavobacteriaceae family in the saliva. As mentioned above, the *Capnocytophaga* (dominant members of the Flavobacteriaceae in these data) are highly metabolically dependent, and positive correlations among organisms are enriched in oral biofilm associated organisms generally (see Figures 3.3 and 3.4).

In addition to these on-diagonal outliers (Figure 3.6), several additional groups of organisms lay off the trend of functional and phylogenetic similarity. That is, some co-occurring/co-exclusive microbes were evolutionarily distant but functionally more similar than expected (below trend), while others were evolutionarily close but functionally distinct (above trend). Several relationships in the upper-left represent single functionally diverse clades that are also widely co-distributed among bilateral or related body sites, such as the Actinobacteria (skin) and Fusobacteriaceae (oral). Such clades' functional distances reflect a relatively high level of within-clade diversity, raising the possibility that a combination of environmental perturbation with a highly structured microenvironment might help to maintain a tension of high functional diversity within a limited phylogenetic range.

In the oral cavity, co-occurrence of such outliers, appearing off the trend of functional and phylogenetic similarity (Figure 3.6) was limited to low-abundance community members, with some exceptions. Abundant signature taxa such as *Streptococcus* and *Neisseria* often excluded clades with more stringent environmental oxygen requirements regardless of their specific degree of relatedness. *Prevotella* - evolutionarily distant from the *Neisseria* but sharing

much functional potential as defined by orthologous gene clusters - exhibited a negative association in the tonsil. Because of their functional similarity, particularly their shared metabolic requirements (both with varying degrees of saccharolytic and proteolytic activities at the species level [78,79]), this strongly suggests a co-exclusion due to competition for resources, again in addition to their environmental oxygen requirement. At the family level, Pasteurellaceae (composed of *Actinobacillus, Aggregatibacter, Haemophilus,* and *Pasteurella*) negatively correlate with several other members of the microbiota. Co-occurrence is more common with pairs of functionally similar microbes able to co-exist through combinations of complementarity, commensalism, and cross-feeding of vitamins, amino acids, and other cofactors. Here, the *Prevotella* produce hydrogen, which influences the growth of *Campylobacteria* [80]. *Prevotella* can also be supported by glycine and pyruvate produced from glutathione by *Treponema* species in the periodontal pocket. The most extreme case of organisms both related and correlated were the *Bacteroides* and *Parabacteroides* in the anterior nares. While these clades are not always well-resolved [81], this trend occurred in nine distinct samples (of 282 total), the only ones in which *Bacteroides* occurred nasally at >5% abundance. While a trivial explanation might be misclassification of a small portion of *Bacteroides*, this trend might instead suggest co-occurrence in a metabolic niche that rarely favors either organism but, in the rare occasion of favoring one, permits both.

Discussion

We analyzed ecological interactions among bacteria in the human microbiome using 16S marker gene abundance data from the Human Microbiome Project. Our methods for building a

microbiome-wide microbial association network combined two complementary approaches: an ensemble of similarity/dissimilarity measures and a compendium of generalized boosted linear models. Relationship significance was assessed using a novel nonparametric approach to compositional data analysis, resulting in a network of co-occurrence and co-exclusion relationships representing potential microbial interactions and incompatibilities within and across body sites. Analysis of the network demonstrated strong organization of the human microbiota into body area niches, mostly among closely related individual body sites representing microbial habitats. A few "hub" microbes were observed to act as signature taxa driving the composition of each microcommunity. Many of these were also the dominant species within a body area, for example *Streptococcus* in the oral cavity and *Bacteroides* in the gut, and these highly abundant taxa also frequently co-occurred as connectors among multiple related body sites. *In vivo* mechanisms were available from prior work for many of these associations, and more generally the phylogenetic and functional relatedness of pairs of co-occurring microbes often explained their associations. In particular, phylogenetically related microbes tended to co-occur at proximal or environmentally similar body sites, while distantly related microbes with shared functional capacities tended to compete.

This microbial association network was described from observational data, and the mechanisms underlying any of these putative interactions may be quite diverse. Positive co-occurrence association types could include nutritional cross-feeding, co-aggregation, co-colonization, signaling pathways, and co-survival in similar environments [4,82]. Negative exclusion interactions likewise might span toxin or small molecule production, environmental

modification (to the detriment of microbial neighbors), immunomodulation, or gross overpopulation of a niche. Ecologically, these data alone do not resolve variations of mutualism, commensalism, amensalism, or predator-prey relationships [4,82]. Further, all of these ecological relationships, detected here based on microbial abundance patterns across many subjects, are themselves distinct from the biogeographical "co-occurrence" patterns observed by previous studies of individual microbes within subjects [33,34]. To distinguish between these, future work could include perturbation experiments (e.g. the removal of a species from a defined habitat such as the gut of a gnotobiotic mouse), as these are becoming less difficult to sustain technically [83]. Analytic refinements might instead include defining directionality of relationships in higher-resolution (e.g. temporal) data; for instance, we expect a strict mutualistic relationship (where both partners cannot exist without the other) to be symmetric, whereas the relationship between a prey and a specialized predator is expected to be asymmetric (the prey can occur without its predator, but not vice versa). Negative co-exclusions may have fewer possible initial interpretations, comprising the types of competition outlined above, or they may indicate different, exclusive microbial community states occurring temporally or as linked to host environment [15,54,77].

Methodologically, it is again important to emphasize that detecting significant co-occurrences among members of a population assayed as relative abundances can be surprisingly difficult due to compositionality [29]. That is, an absolute increase in one organism's abundance can result in an apparent relative decrease of all other abundances, leading to spurious correlations. Extensive prior work has explained the problem in microbial

and macroecological settings [53,84], and we have mitigated potential issues in these data through our ensemble approach and by principled calculation of significance thresholds using null distributions that incorporate the degree of similarity due solely to compositional effects (see Methods, Appendix 3). GBLMs were the most distinct method included in this ensemble and share some similarities with recently proposed genetic regulatory network (GRN) reconstruction techniques [85]. GBLMs do provide methodology for discovering GRN-like higher-order interactions in microbial communities, but the accuracy needed to overcome the associated multiple hypothesis testing problems is not yet achievable from available 16S data [86,87]. We anticipate that future studies with species- or strain-level classification of deep shotgun metagenomic sequences may provide sufficient resolution for more detailed networks including such cooperative microbial associations.

While it might be hoped that easily sampled microbiomes such as the saliva would serve as proxies for e.g. the broader oral microbiome, these results suggest that this is not generally the case. There are few strong correspondences among organisms even at closely related body sites, let alone distal sites, and very few cases where microbial abundance is quantitatively predictable from a proxy sample. In the HMP, this may be a feature of a healthy population, and additional relationships (or disruption of existing ones) might emerge in the presence of disease. Environmental factors that strongly impact the healthy microbiome may additionally not be captured for this population (e.g. diet) and can be further investigated by targeted methodology in future cohorts.

This catalog of microbial co-occurrence and co-exclusion relationships thus provides an initial glimpse of potential mechanisms of community organization throughout the human microbiome. While this computational methodology can be applied to any communities assayed using marker gene sequencing, it is interesting to conclude by noting that the resolution of the resulting network is limited by the specificity of 16S sequence binning. The network discussed here, for example, leverages two specific hypervariable regions for taxonomic classification, each with strengths and weaknesses, and neither individually adequate for sequence classification at the species level [88,89]. Since it is likely that additional microbial associations will occur at the species or strain level, we anticipate that further community structure will emerge during analysis of metagenomic shotgun sequences taxonomically binned at a finer level of detail. Community shotgun sequencing will also provide functional information regarding metabolism, signaling, and, again, potential physical mechanisms of interaction, which can in turn be matched against complete reference genomes for co-occurring strains. Perturbation analyses in co-culture or, eventually, longitudinal studies in human cohorts will provide an intriguing means of investigating the impact of these microbial "wiring" diagrams on human health.

Chapter 3 Bibliography

1. Saffo MB (1993) Coming to Terms with a Field - Words and Concepts in Symbiosis (Vol 14, Pg 29, 1993). Symbiosis 15: 181-181.

2. William Z. Lidicker J (1979) A Clarification of Interactions in Ecological Systems. BioScience 29: 475-477.

3. Diamond JM (1975) Assembly of species communities. In: Cody ML, Diamond JM, editors. Ecology and Evolution of Communities. Cambridge, MA: Harvard University Press. pp. 342-444.

4. Konopka A (2009) What is microbial community ecology? The ISME Journal 3: 1223-1230.

5. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, et al. (2006) Symbiotic insights through metagenomic analysis of a microbial consortium. Nature 443: 950-955.

6. Marx CJ (2009) Microbiology. Getting in touch with your friends. Science 324: 1150-1151.

7. Follows MJ, Dutkiewicz S, Grant S, Chisholm SW (2007) Emergent biogeography of microbial communities in a model ocean. Science 315: 1843-1846.

8. Freilich S, Kreimer A, Meilijson I, Gophna U, Sharan R, et al. (2010) The large-scale organization of the bacterial network of ecological co-occurrence interactions. Nucleic Acids Res 38: 3857-3868.

9. Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, et al. (2011) Marine bacterial, archaeal and protistan association networks reveal ecological linkages. The ISME Journal 5: 1414-1425.

10. Simon C, Daniel R (2011) Metagenomic analyses: past and future trends. Appl Environ Microbiol 77: 1153-1161.

11. Parfrey LW, Walters WA, Knight R (2011) Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. Front Microbiol 2: 153.

12. Sansonetti PJ (2004) War and peace at mucosal surfaces. Nat Rev Immunol 4: 953-964.

13. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE (2005) Defining the normal bacterial flora of the oral cavity. J Clin Microbiol 43: 5721-5732.

14. Zaura E, Keijser BJ, Huse SM, Crielaard W (2009) Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol 9: 259.

15. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature.

16. Frank DN, Zhu W, Sartor RB, Li E (2011) Investigating the biological and clinical significance of human dysbioses. Trends Microbiol 19: 427-434.

17. Kinross JM, Darzi AW, Nicholson JK (2011) Gut microbiome-host interactions in health and disease. Genome Medicine 3: 14.

18. Reid G, Younes JA, Mei HCVd, Gloor GB, Knight R, et al. (2011) Microbiota restoration: natural and supplemented recovery of human microbial communities. Nature Reviews Microbiology 9: 27-38.

19. Horner-Devine MC, Silver JM, Leibold MA, Bohannan BJ, Colwell RK, et al. (2007) A comparison of taxon co-occurrence patterns for macro- and microorganisms. Ecology 88: 1345-1353.

20. Chaffron S, Rehrauer H, Pernthaler J, von Mering C (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res 20: 947-959.

21. Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, et al. (2011) Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J 5: 1414-1425.

22. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, et al. (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443: 950-955.

23. Hojo K, Nagaoka S, Ohshima T, Maeda N (2009) Bacterial interactions in dental biofilm development. J Dent Res 88: 982-990.

24. Huang R, Li M, Gregory RL (2011) Bacterial interactions in dental biofilm. Virulence 2: 435-444.

25. Kuramitsu HK, He X, Lux R, Anderson MH, Shi W (2007) Interspecies interactions within oral microbial communities. Microbiol Mol Biol Rev 71: 653-670.

26. Silverman MS, Davis I, Pillai DR (2010) Success of self-administered home fecal transplantation for chronic Clostridium difficile infection. Clin Gastroenterol Hepatol 8: 471-473.

27. Legendre P, Legendre L (1983) Numerical ecology. Amsterdam: Elsevier Science B.V.

28. Gross EL, Leys EJ, Gasparovich SR, Firestone ND, Schwartzbaum JA, et al. (2010) Bacterial 16S sequence analysis of severe caries in young permanent teeth. J Clin Microbiol 48: 4121-4128.

29. Aitchison J (1981) A New Approach to Null Correlations of Proportions. Mathematical Geology 13: 175-189.

30. Aitchison J (1982) The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society Series B (Methodological) 44: 139-177.

31. Aitchison J. A Concise Guide to Compositional Data Analysis; 2003; Girona, Spain.

32. The Human Microbiome Consortium (in review) A framework for human microbiome research.

33. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, et al. (2009) The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med 1: 6ra14.

34. Stearns JC, Lynch MD, Senadheera DB, Tenenbaum HC, Goldberg MB, et al. (2011) Bacterial biogeography of the human digestive tract. Sci Rep 1: 170.

35. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.

36. Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One 6: e27310.

37. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141-145.

38. Buehlmann P (2006) Boosting for high-dimensional linear models. The Annals of Statistics 34: 559-583.

39. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Methodological 58: 267-288.

40. Kuhn M (2008) Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28: 1-26.

41. Hothorn T, Buhlmann P, Kneib T, Schmid M, Hofner B (2010) Model-based Boosting 2.0. Journal of Machine Learning Research 11: 2109-2113.

42. Knights D, Kuczynski J, Koren O, Ley RE, Field D, et al. (2011) Supervised classification of microbiota mitigates mislabeling errors. The ISME journal 5: 570-573.

43. Date SV, Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nature Biotechnology 21: 1055-1062.

44. Sarkar SK, Chang CK (1997) The Simes method for multiple hypothesis testing with positively dependent test statistics. Journal of the American Statistical Association 92: 1601-1608.

45. Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys 70: 066111.

46. van Dongen S (2008) Graph Clustering Via a Discrete Uncoupling Process. SIAM J on Matrix Analysis and Application 30: 121-141.

47. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

48. Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE 5: e9490.

49. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22-28.

50. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, et al. (2011) Moving pictures of the human microbiome. Genome Biol 12: R50.

51. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, et al. (in review) Composition of the Adult Digestive Tract Microbiome Based on Seven Mouth Surfaces, Tonsils, Throat and Stool Samples.

52. The Human Microbiome Consortium (in review) Structure, Function and Diversity of the Human Microbiome in an Adult Reference Population.

53. Friedman J, Alm E (in review) Severe Artifacts Prevent Clustering and Correlation of Genomic Survey Data.

54. Ravel J, Gajer P, Abdob Z, Schneider GM, Koenig SSK, et al. (2011) Vaginal microbiome of reproductive-age women. PNAS 108: 4680-4687.

55. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, et al. (2009) Topographical and temporal diversity of the human skin microbiome. Science 324: 1190-1192.

56. Nadell CD, Foster KR, Xavier JB (2010) Emergence of spatial structure in cell groups and the evolution of cooperation. PLoS Comput Biol 6: e1000716.

57. Barabási A-L, Albert R (1999) Emergence of Scaling in Random Networks. Science 286: 509-512.

58. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440-442.

59. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical Organization of Modularity in Metabolic Networks. Science 297: 1551-1555.

60. Kingsley VV, Hoeniger JF (1973) Growth, structure, and classification of Selenomonas. Microbiol Mol Biol Rev 37: 479-521.

61. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, et al. (2009) Bacterial community variation in human body habitats across space and time. Science 326: 1694-1697.

62. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480-484.

63. Naito M, Hirakawa H, Yamashita A, Ohara N, Shoji M, et al. (2008) Determination of the genome sequence of Porphyromonas gingivalis strain ATCC 33277 and genomic comparison with strain W83 revealed extensive genome rearrangements in P. gingivalis. DNA Res 15: 215-225.

64. Kolenbrander PE, Parrish KD, Andersen RN, Greenberg EP (1995) Intergeneric coaggregation of oral Treponema spp. with Fusobacterium spp. and intrageneric coaggregation among Fusobacterium spp. Infect Immun 63: 4584-4588.

65. Kolenbrander PE, Andersen RN (1989) Inhibition of coaggregation between Fusobacterium nucleatum and Porphyromonas (Bacteroides) gingivalis by lactose and related sugars. Infect Immun 57: 3204-3209.

66. Kolenbrander PE, Palmer RJ, Jr., Periasamy S, Jakubovics NS (2010) Oral multispecies biofilm development and the key role of cell-cell distance. Nat Rev Microbiol 8: 471-480.

67. Bradshaw DJ, Marsh PD, Watson GK, Allison C (1998) Role of Fusobacterium nucleatum and coaggregation in anaerobe survival in planktonic and biofilm oral microbial communities during aeration. Infect Immun 66: 4729-4732.

68. Carlsson J, Iwami Y, Yamada T (1983) Hydrogen peroxide excretion by oral streptococci and effect of lactoperoxidase-thiocyanate-hydrogen peroxide. Infect Immun 40: 70-80.

69. Tanner ACR, Listgarten MA, Ebersole JL, Strezempko MN (1986) Bacteroides-Forsythus Sp-Nov, a Slow-Growing, Fusiform Bacteroides Sp from the Human Oral Cavity. International Journal of Systematic Bacteriology 36: 213-221.

70. Downes J, Vartoukian SR, Dewhirst FE, Izard J, Chen T, et al. (2009) Pyramidobacter piscolens gen. nov., sp. nov., a member of the phylum 'Synergistetes' isolated from the human oral cavity. Int J Syst Evol Microbiol 59: 972-980.

71. Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, et al. (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. PNAS 104: 11889-11894.

72. Walter J, Ley R (2011) The Human Gut Microbiome: Ecology and Recent Evolutionary Changes. Annu Rev Microbiol 65: 411-429.

73. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031.

74. Murgas Torrazza R, Neu J (2011) The developing intestinal microbiome and its relationship to health and disease in the neonate. J Perinatol 31 Suppl 1: S29-34.

75. Zaric S, Bojic B, Jankovic L, Dapcevic B, Popovic B, et al. (2009) Periodontal therapy improves gastric Helicobacter pylori eradication. Journal of dental research 88: 946-950.

76. Dunny GM, Brickman TJ, Dworkin M (2008) Multicellular behavior in bacteria: communication, cooperation, competition and cheating. Bioessays 30: 296-298.

77. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105-108.

78. Ryan KJ, Ray CG, Sherris JC (2004) Sherris medical microbiology : an introduction to infectious diseases. New York: McGraw-Hill.

79. Shah HN, Collins DM (1990) Prevotella, a new genus to include Bacteroides melaninogenicus and related species formerly classified in the genus Bacteroides. International Journal of Systematic Bacteriology 40: 205-208.

80. Wyszynska A, Tomczyk K, Jagusztyn-Krynicka EK (2007) Comparison of the localization and post-translational modification of Campylobacter coli CjaC and its homolog from Campylobacter jejuni, Cj0734c/HisJ. Acta Biochim Pol 54: 143-150.

81. Karlsson FH, Ussery DW, Nielsen J, Nookaew I (2011) A closer look at bacteroides: phylogenetic relationship and genomic implications of a life in the human gut. Microbial ecology 61: 473-485.

82. Raes J, Bork P (2008) Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol 6: 693-699.

83. Faith JJ, Rey FE, O'Donnell D, Karlsson M, McNulty NP, et al. (2010) Creating and characterizing communities of human gut microbes in gnotobiotic mice. ISME J 4: 1094-1098.

84. Jackson D (1997) Compositional data in community ecology: the paradigm or peril of proportions? Ecology 78: 929-940.

85. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5: e12776.

86. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194-2200.

87. Schloss PD, Gevers D, Westcott SL (in press) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE.

88. Chakravorty S, Helb D, Burday M, Connell N, Alland D (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods 69: 330-339.

89. Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, et al. (2010) Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. World J Gastroenterol 16: 4135-4144.

Chapter 4: Integrative Genomics of Sexual Dimorphism in COPD

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is the third leading cause of death in the United States. Epidemiologic studies report more women than men dying from COPD with additional data suggesting that females may be biologically more susceptible to the disease. To identify molecular pathways associated with sexually dimorphic features of COPD, we used lung tissue gene expression data from the Lung Genomics Research Consortium to identify 959 genes with sexually dimorphic patterns of differential expression in the presence of COPD ("sexually dimorphic and COPD differential" or "SDCD" genes). Gene set enrichment analysis highlighted processes including chemotaxis, inflammatory responses, cell morphogenesis, and VEGF signaling. We observed that a subset of SDCD genes are more likely to be targeted by progesterone, vitamin D, and PPAR-gamma, and their promoters contain an overrepresentation of estrogen and androgen response elements. We also observed sex-specific differential methylation for one third of the SDCD genes. A sex-stratified eQTL analysis identified 94 SDCD genes with suggested genetic regulation of gene expression. Our study represents the first systematic integrative genomic survey of sexual dimorphic gene expression in COPD, illustrates involvement of key gene regulation mechanisms such as hormones, methylation, and genetic modification, and highlights the importance of sex-specific approaches to the diagnosis, treatment, and primary prevention of COPD.

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is the third leading cause of death in the US [1,2]; for over a decade, more women than men have died of COPD [3,4]. Increases in cigarette smoking by women in the 1960s and 1970s may account for some of the observed increase in female mortality. However, even when controlling for smoking history, females are more susceptible to COPD than men [5,6], develop more severe symptoms at an earlier age [7,8], and have worse health outcomes [9]. In addition to variable susceptibility and severity of COPD, the disease sub-phenotypes differ between the sexes, with chronic bronchitis more prevalent in females and emphysema more prevalent in males [9,10].

Multiple mechanisms may contribute to the sexually dimorphic features of COPD, including variable xenobiotic metabolism of tobacco smoke, differential airway inflammation and bronchial hyperresponsiveness to noxious stimuli, and differences in lung anatomy and size [5,6]. In addition, sex hormones likely influence airway function, response to injury and repair, as reflected in sex differences in age-dependent risk of childhood asthma [11], the larger decrease in lung function in adolescent girls compared to adolescent boys who smoked cigarettes [12], and the increased risk for COPD among post-menopausal women [13]. Estrogen has also been associated with increased metabolism of nicotine through cytochrome P450 (CYP) without increased activity of the detoxifying enzymes [5], producing higher oxidative stress in the lungs of female smokers. Reduced endogenous testosterone levels have been observed in men with COPD [14,15] and testosterone supplementation has been investigated to improve muscle strength in men with COPD [16,17]. In addition, vitamin D receptor [18,19] and peroxisome proliferator-activated receptor gamma (PPARG) [20] have been implicated in lung

disease and manifest sex specific functions [21-23] and may contribute to sex differences in COPD.

While many lines of evidence indicate sex differences in COPD biology and physiology [5,24,25], there is little understanding of the molecular and cellular mechanisms of COPD sexual dimorphism. Gene expression studies have identified both genes and pathways relevant to COPD [26-29], but none of these studies has systematically interrogated sex-specific associations. The Lung Genomics Research Consortium (LGRC; www.lung-genomics.org) provides genetic, molecular, and quantitative phenotype data for patient samples in the NHLBI's Lung Tissue Research Consortium (LTRC) biorepository. The LGRC data include gene expression, DNA methylation, and genotype information on several hundred COPD and control lung tissue samples. Here we describe an analysis of the LGRC data to identify molecular processes and pathways that differ in lung tissue from male and female subjects. We identified 959 genes with sexually dimorphic differential expression that represent pathways including chemotaxis, inflammatory response, cell morphogenesis, and VEGF signaling. These 959 genes are enriched for hormone response elements, and we have also found evidence for sexually dimorphic methylation and genetic regulation, suggesting potential mechanisms regulating sexually dimorphic gene expression.

Methods

Reproducible research

We have provided the code used in this analysis as a Bioconductor package COPDSexualDimorphism

(http://www.bioconductor.org/packages/release/bioc/html/COPDSexualDimorphism.html) and a companion experiment data package COPDSexualdimorphism.data.

LGRC data processing

We retrieved clinical metadata, gene expression, and methylation data from the LGRC Data Portal (https://www.lung-genomics.org/research); genotype data have been deposited in the dbGaP repository (accession number: phs000624.v1.p1 in progress). The data were processed, all with Human Genome Version 19 (hg19) annotation, as follows.

LGRC samples

There are 1,359 total lung tissue and blood samples in the LGRC, comprising 576 samples from interstitial lung disease (ILD) patients, 511 from COPD patients, and 272 from controls (mostly normal lung tissues from COPD/ILD-free patients with lung cancer). We first removed 708 blood samples, leaving 651 whole lung samples. Of the whole lung samples, 283 clinically diagnosed ILD samples and 36 samples with pathological features of ILD were removed. To reduce genetic heterogeneity of the data, 16 non-Caucasian samples were removed. Eight additional samples were removed because they were labeled as COPD but had spirometric measures in the normal range. Seven more samples were removed because they were labeled as former smokers but had zero pack years of cigarette smoking in the record. We also removed 15 samples whose smoking statuses or pack years were unknown. Two samples were excluded because they had high pre-bronchodilator FEV1/FVC ratios (1.3 and 2.4), and an additional 15 samples were removed based on the clinical diagnoses. Some subjects contributed more than one sample (for example samples from left and right lungs); we removed 15

duplicate samples and kept the samples with more complete genetic and molecular phenotypes. Finally, we flagged 26 control samples with low diffusion capacity (DLCO < 80) to perform sensitivity analysis. In the end 179 COPD samples (102 male, 77 female) and 75 control samples (31 male, 44 female) passed these criteria (254 total) and were used in the subsequent analyses.

Gene expression data processing

GCRMA-normalized gene expression data from Agilent Whole Human Genome 4 × 44K arrays (G4112F, Agilent Technologies) were downloaded from LGRC Data Portal (www.lung-genomics.org/research; accessed November 2011). We mapped probe names to Ensembl gene ID using BiomaRt R package. Only one of the probes with highest variance among the probes mapped to the same Ensembl gene ID was kept (14,497 probes left). Surrogate variable analysis [30] identified no significant batch effect.

SDCD expression analysis

To identify genes with sexually dimorphic differential expression, we performed two complementary analyses: stratification by sex and stratification by COPD-control status. Here we describe the analysis for case-control stratified analysis, and the sex-stratified analysis proceeds in the same manner. For each Ensembl gene in the case stratum (N=164), we used the limma R package to fit a linear model using the log expression level as outcome and sex indicator as predictor, adjusting for age and pack years of smoking (Model 4). The same models were fit for the control stratum (N=65, Model 3).

$$\text{Gene Expression} \mid \text{control} = \beta_{0,\text{control}} + \beta_{1,\text{control}}\,\text{Sex} + \beta_{2,\text{control}}\,\text{Pkyr} + \beta_{3,\text{control}}\,\text{Age} + \varepsilon_{\text{control}} \quad (3)$$

$$\text{Gene Expression} \mid \text{COPD} = \beta_{0,\text{COPD}} + \beta_{1,\text{COPD}}\,\text{Sex} + \beta_{2,\text{COPD}}\,\text{Pkyr} + \beta_{3,\text{COPD}}\,\text{Age} + \varepsilon_{\text{COPD}} \quad (4)$$

The standard errors of the coefficient associated with sex indicator, $SE(\beta_{1,\text{COPD}})$ and $SE(\beta_{1,\text{control}})$, were combined, and the difference between the coefficients were converted to Cohen's d and evaluated with a two-sided z-test: $p_{\text{COPD-control}} = 2(1 - \Phi(|\beta_{\text{COPD}} - \beta_{\text{control}}|/S_{\text{pooled}}))$ [31,32]. Benjamini-Hochberg FDR correction was applied to the p-values. Similar analysis was performed for sex-stratified data, and the genes with both adjusted p-values less than 0.25 were selected as SDCD genes.

Validation of SDCD genes with independent datasets

Gene expression profiles were downloaded from Gene Expression Omnibus (GEO) through the R package GEOquery (accession numbers: GSE8581 for Bhattacharya [33] and GSE37147 for Steiling [28]). We performed quantile normalization using affyPLM R package and batch effect correction using the SVA package. We adjusted for 7 and 24 significant surrogate variables detected in Bhattacharya and Steiling datasets, respectively.

The Bhattacharya dataset uses Affymetrix Human Genome U133 Plus 2.0 arrays (HG-U133_Plus_2), which has a total of 54,675 probes. 1,867 probes were mapped to SDCD genes through Ensembl gene identifiers, which were used as unique identification for SDCD genes, representing 942 of the 959 SDCD genes. The same regression models as described in "SDCD expression analysis" section above were fit to the data except only age, not pack years, was adjusted due to the limited data available publicly. As in the "SDCD analysis," the FDR threshold of 0.25 was used in each stratified analysis before combining the results by set intersection. The Steiling's dataset used the Affymetrix GeneChip Human Gene 1.0 ST arrays (HuGene10stv1_Hs_ENSG), which has 19,793 probes. Of those probes, 948 were mapped to the

959 SDCD genes through Entrez and Ensembl gene identifiers, representing 944 SDCD genes. SDCD regression models were fit, adjusted for age and pack year of smoking; FDR cutoff of 0.25 was applied; results from two stratified models were combined.

Functional enrichment analysis

Functional and pathway enrichment analyses were carried out by various methods, namely a conditional hypergeometric test by GOstats R package, a gene set enrichment analysis (GSEA) by GAGE R package, and a Fisher's exact test by Ingenuity Pathway Analysis (IPA, Ingenuity Systems, www.ingenuity.com). Genes presented in the expression profile were mapped to Entrez ids via BiomaRt R package. For the GOstats approach, all of the mapped Entrez genes were used as the background set, and the Entrez ids corresponding to SDCD genes were tested for enrichment in the three classes of GO terms. For the GSEA approach, two ranking approaches were used. First all available genes were ranked by the absolute difference between the coefficients $|\Delta\beta_{\text{COPD-control}}|$ and secondly by the adjusted p-value ($p_{\text{COPD-control}}$). For canonical pathway analysis, the Ensembl gene ids of the SDCD genes were imported into IPA (Ingenuity Systems, www.ingenuity.com), and we used the IPA canonical analysis. For all of the approaches except GSEA, we selected functions with enrichment p-value < 0.05 and at least three genes overlap.

Functional network clustering

For each pair of functional annotation terms the number of shared SDCD genes and the Jaccard index was calculated. We constructed a network with functional terms as nodes and the shared genes as edges and filtered for edges with Jaccard index greater than or equal to 0.2.

Edges connecting nodes representing GO terms that are directly related through parent-progeny relationship in the GO hierarchical structure were removed. The networks were visualized using Cytoscape [34]. We used a weighted force-directed layout method [35] to visualize the networks and used GLay community structure analysis (fast-greedy algorithm) [36] to identify functional clusters.

Differential methylation analysis

Methylation data processing

Normalized, quality-controlled, and batch effect-corrected percent methylation data from comprehensive high-throughput arrays for relative methylation (CHARM) [37] for 178 samples (134 COPD, 44 controls) were downloaded from LGRC Data Portal (www.lung-genomics.org/research; accessed January 2013). Pre-processing of CHARM methylation arrays used the CHARM Bioconductor Package to perform standard CHARM quality control, normalization, and smoothing to produce percent methylation at each of the 2,162,405 probes. Probes were mapped to positions in Human Genome version 19 (hg19). Surrogate Variable Analysis (SVA) [30] was used to detect batch effect in the logit-transformed percent methylation. Two significant surrogate variables, corresponding to the first two principle components, were removed using the R Package SWAMP (cran.r-project.org/web/packages/swamp/), and the batch-corrected values were then transformed back into percent methylation. Only data from COPD and control samples were used for the differential methylation analysis.

We defined variably methylated regions (VMRs) as previously described [38,39]. Briefly, we first calculated for all probes the median absolute deviation (MAD) of percent methylation across all samples in our study. Adjacent probes with high MAD (>80th percentile) within 300bp were clustered into VMRs. We required at least five probes in a VMR and identified 12,094 VMRs.

Sexually dimorphic and differential methylation

The SDCD methylation analysis described above was performed with logit-transformed percent methylation of each VMRs as outcome. Starting with 12,094 VMRs, we used GenomicRanges R package to search for SDCD genes within 10kb distance on either side of the mid points of the VMRs. We found 888 VMRs with at least one SDCD gene nearby. These accounted for 395 SDCD genes. We then ran SDCD analysis on these 888 VMRs with the linear model adjusting for pack years and age, same as SDCD expression analysis. Because sex-stratified analysis yielded no significant VMR at FDR cutoff 0.05, we proceeded with only COPD-stratified analysis, which identified 387 SDCD VMRs at FDR cutoff of 0.05. A VMR is associated with a SDCD gene if the mid point of the VMR is within 10kb distance to the gene's transcription start site.

Nuclear receptor enrichment analysis

Cistrome database

Lists of targets of estrogen (ESR1/2), androgen (AR), progesterone (PGR), vitamin D (VDR), and PPAR-gamma (PPARG) were downloaded from Cistrome database (cistrome.org, accessed 1/28/2013). For each of the nuclear receptors, Cistrome assigns to each gene a rank

product score, which measures the strength of regulatory potential between the nuclear receptor and the gene and can be interpreted as a p-value [40]. Thus, we defined significant regulatory targets by selecting genes with rank product cutoff of 0.25 (see justification of the cutoff in Appendix 4A). For each of the regulatory target list, we used Fisher's exact test, with genes in the expression probe set as background, to assess enrichment.

JASPAR motif scan

To create our motif prior, we downloaded the position weight matrixes (PWM) of ESR1, ESR2, and AR motifs from the JASPAR database. To search for motif target candidates, the motif score of each candidate S was defined as motif score = $\log[P(S|M)/P(S|B)]$, where $P(S|M)$ is the probability to observe sequence S given the motif M, and $P(S|B)$ is the probability to observe sequence S given the genome background B. To define motif targets, we modeled the motif score distribution by randomly sampling the genome 10^6 times. Targets of motifs were then defined as those with a score at a significance level of $p<10^{-5}$. We associated genes with these motif targets if that target fell within its promoter region [750bp, +250bp]. We also used Fisher's exact test, with genes in the expression probe set as background, to assess enrichment.

Other databases

A list of estrogen response elements (ERE) was downloaded from the mouse and human ERE database (www.mapageweb.umontreal.ca/maders/eredatabase, accessed 1/17/2013). 2,980 of these genes were present in our gene expression probe set, and 217 SDCD genes contained at least one ERE. A list of androgen responsive genes (ARG) was downloaded form ARGDB (argdb.fudan.edu.cn, accessed 1/22/2013). 1,344 of the ARGs were in our gene expression probe

set, and 103 were SDCD genes. Like Cistrome and JASPAR, we used a Fisher's exact test to assess enrichment of the genes with hormone response elements.

Sex-specific eQTL analysis

Genotype data processing for eQTL analysis

Genotyping of 535 samples (219 ILD, 224 COPD, 92 controls) were from Illumina Infinium HD Assay with Human Omni 1 QuAd and Human Omni 2.5 QuAd arrays. Genotype calling and processing were performed at University of Colorado Boulder where the genotyping was done. Genome Studio (default settings) was used to call genotypes and genotypes from Omni 1 arrays were imputed to the Omni 2.5 array design and processed together thereafter (2,443,179 markers). We converted the genotype files into PLINK formats (.bim and .fam) and used PLINK to perform the following quality control steps. We considered only 171 COPD and control samples with both genotypes and gene expression.

For SNP quality controls, 90 male and 81 female samples were analyzed separately. In males and females, respectively, 322 and 327 markers were excluded based on Hardy-Weinberg Equilibrium test ($p \leq 0.001$); 37,110 and 34,968 markers were excluded because they were missing in more than 10% of the samples; and 1,119,893 and 1,141,434 SNPs were excluded because the minor allele frequencies were less than 0.05. In sum, 1,232,344 SNPs remained after the QC steps (1,300,303 male and 1,279,600 female). No sample was excluded based on the genotyping rate (call rates > 0.95).

We removed 7 samples with discordance between genetic sex and reported sex. Inbreeding coefficient *F* was calculated for each subject after SNP pruning and Linkage Disequilibrium

(LD) filtering, as an extreme (especially negative) F reflects excess in heterozygosity and homozygosity. All samples passed the standard cutoff of $|F| > 0.2$. Although subjects in the study are reported to be unrelated, we used identity by descent (IBD) to identify and remove one sample from a subject whose both left and right lungs were genotyped, keeping the sample with more complete molecular data. In the end, we had 126 COPD and 37 control samples for the sex-stratified eQTL analysis (86 males, 77 females, 163 total).

Sex-stratified eQTL analysis

We performed genome-wide expression quantitative trait loci (eQTL) analysis, which fits mRNA expression to the following model:

$$Gene\ Expression = \beta_0 + \beta_1 ADD + \beta_2 Pkyr + \beta_3 Age + \varepsilon$$

where ADD is the allele dosage of a genotype (coded as 0, 1, and 2 for the number of minor alleles). Only *cis*-variants (SNPs within 100kb upstream and 10kb downstream from TSSs) of SDCD genes were considered. We used PLINK [41] to perform the eQTL analyses.

Similar to the SDCD gene expression analysis, models for males and females were fit separately and the allelic effects (β_1) were contrasted. Benjamini-Hochberg multiple hypothesis correction was performed. We selected eQTL with FDR < 0.05 and more than five data points with homozygous minor alleles.

Results

Lung Genomics Research Consortium (LGRC) cohort

We selected data from 254 subjects from the LGRC (www.lung-genomics.org) (179 COPD, 75 controls) (see Methods). Clinical characteristics are shown in Table 1. Male and female subjects have different distributions of COPD severity; 36.4% of females were classified in GOLD stage 4 (very severe COPD) compared to 20.6% of the males, although males smoked more than females. Fifteen percent of the subjects (N=37) did not self-report cigarette smoking and these samples were flagged for a sensitivity analyses that focused only on ever-smokers. Female COPD subjects had lower diffusing capacity (DLCO) than male COPD subjects; control samples with low diffusing capacity (DLCO < 80, N=26) were flagged for removal to address potential misclassification bias.

The LGRC includes genotype, gene expression, and methylation data. Of these three data types, gene expression was available for 229 of the 254 samples and methylation for 178 samples. Moreover 163 had genome-wide genotypes in addition to gene expression profiles. The exact breakdown of the number of samples and data type used in our subsequent analyses is given in Table S1, and a summary of data processing is given in Figure S1.

Characteristic	Male-COPD (N=102)	Female-COPD (N=77)	Male-Control (N=31)	Female-Control (N=44)
Age – years	66.55 ± 9.14	64.01 ± 9.74	66.97 ± 11.74	62.18 ± 12.31
Pack years of smoking	60.48 ± 42.41	44.53 ± 25.02	37.45 ± 45.05	10.57 ± 15.53
Smoking status – no. (%)				
Current smoker	4 (3.9)	6 (7.8)	0 (0.0)	1 (2.3)
Former smoker	95 (93.1)	68 (88.3)	22 (71.0)	21 (47.7)
Never smoker*	3 (2.9)	3 (3.9)	9 (29.0)	22 (50.0)
FEV1/FVC	0.48 ± 0.15	0.49 ± 0.15	0.77 ± 0.06	0.79 ± 0.06
DLCO	59.09 ± 22.30	50.47 ± 23.32	85.50 ± 17.56	88.26 ± 13.73
GOLD stage – no. (%)				
0-At Risk	0 (0.0)	0 (0.0)	31 (100.0)	44 (100.0)
1-Mild COPD	11 (10.8)	6 (7.8)	0 (0.0)	0 (0.0)
2-Moderate COPD	52 (51.0)	32 (41.6)	0 (0.0)	0 (0.0)
3-Severe COPD	18 (17.6)	11 (14.3)	0 (0.0)	0 (0.0)
4-Very Severe COPD	21 (20.6)	28 (36.4)	0 (0.0)	0 (0.0)

* flagged for sensitivity analysis.

Table 4.1: Clinical characteristics of LGRC cohort stratified by sex and COPD status

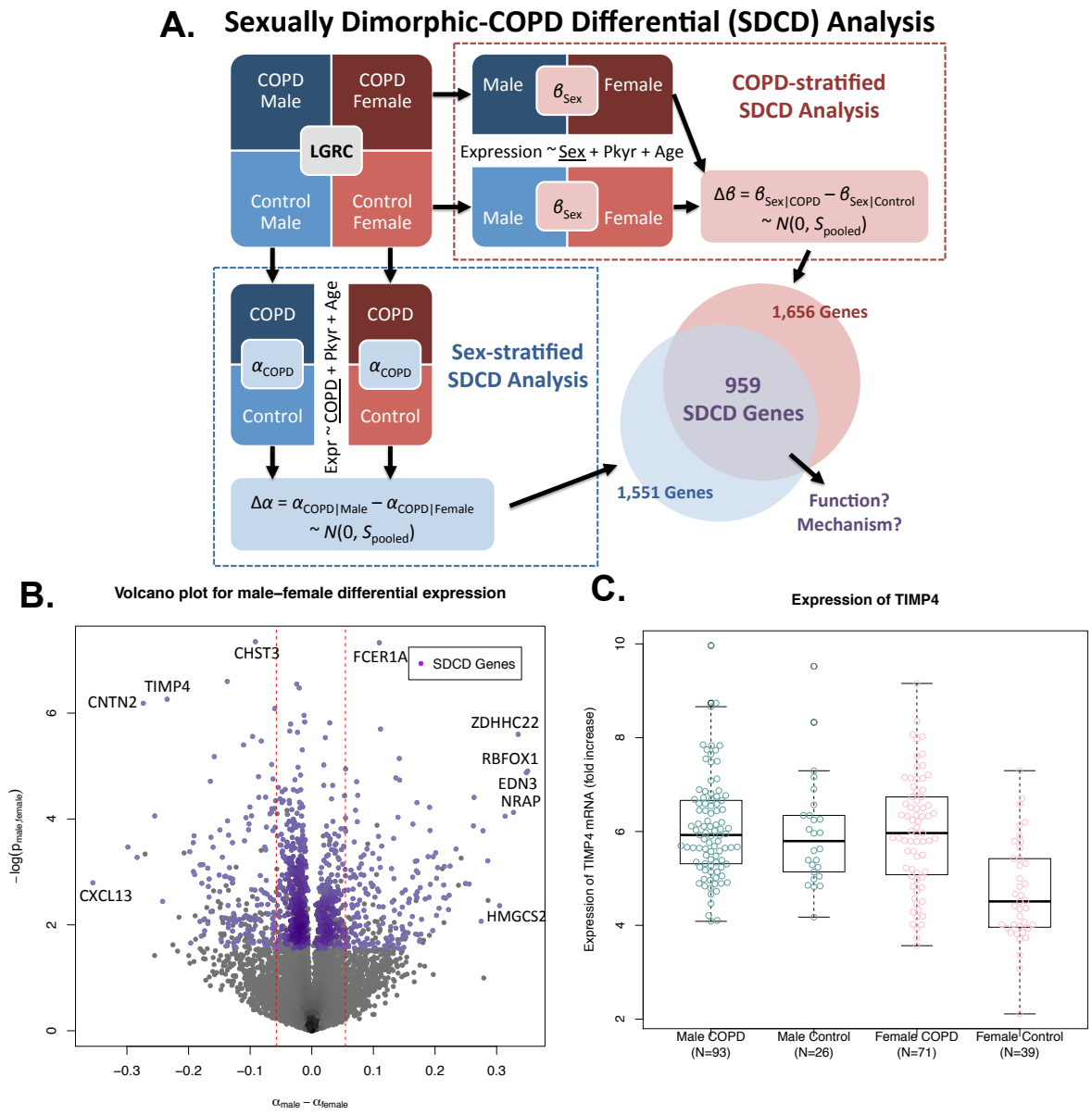


Figure 4.1: An overview of sexually dimorphic and COPD differential (SDCD) analysis. (A) Two parallel analyses were performed: sex-stratified analysis and COPD-stratified analysis. For each of the analyses, the differences between regression coefficients from the two strata were assessed for significance via Cohen's *d*. Finally the genes identified by both approaches were selected and called "SDCD genes." (B) Volcano plot from sex-stratified analysis. The purple points represent SDCD genes; shading represents density of points. Dashed vertical red lines mark the top and bottom 5% quantiles of $\Delta\alpha_{\text{male-female}} = \alpha_{\text{male}} - \alpha_{\text{female}}$. Genes with the largest absolute difference $\Delta\alpha_{\text{male-female}}$ include CXCL13, CTCN2, TIMP4, ZDHHC22, RBFOX1, EDN3, NRAP, and HMGCS2. (C) Expression levels of TIMP4 stratified by sex and COPD status. TIMP4 expression is lowest in female controls. Appearing in the top corner of the volcano plot, TIMP4 expression is unchanged in males with COPD compared to control but more highly expressed in females with COPD compared to control.

Stratified analysis identifies sexually dimorphic and COPD differential (SDCD) genes

To study sexually dimorphic differential expression in COPD, we performed two stratified analyses: stratified by sex and by COPD status. We then combined the results from the two complementary analyses to define a set of "sexually dimorphic and COPD differential" (SDCD) genes. Figure 4.1A shows an outline of our approach. To avoid potential confounding effects from sex-specific regulation of genes on the sex chromosomes, we only considered autosomal genes in all subsequent analyses.

Sex- and COPD-stratified analysis

First we explored COPD-related differential gene expression patterns separately for males and females using linear models, which are effective for the study of sexually dimorphic gene expression [42]. For each gene we calculated a sex-specific regression coefficient representing differential expression in COPD compared to control lung tissues using the following models:

$$\text{Gene Expression} \mid \text{male} = \alpha_{0,\text{male}} + \alpha_{1,\text{male}}\,\text{COPD} + \alpha_{2,\text{male}}\,\text{Pkyr smoked} + \alpha_{3,\text{male}}\,\text{Age} + \varepsilon_{\text{male}} \quad (1)$$

$$\text{Gene Expression} \mid \text{female} = \alpha_{0,\text{female}} + \alpha_{1,\text{female}}\,\text{COPD} + \alpha_{2,\text{female}}\,\text{Pkyr smoked} + \alpha_{3,\text{female}}\,\text{Age} + \varepsilon_{\text{female}} \quad (2)$$

Here Pkyr smoked is total pack-years of cigarette smoking (equal to the number of packs of cigarettes smoked daily multiplied by the number of total years smoked); the coefficient $\alpha_{1,\text{male}}$ represents the degree of expression difference associated with having COPD in males; and $\alpha_{1,\text{female}}$ represents the difference in females. We will hereafter refer to $\alpha_{1,\text{male}}$ as α_{male} and $\alpha_{1,\text{female}}$ as α_{female}.

By stratifying the analysis by sex, we have separate coefficients for males and females, representing sex-specific differential expression patterns in COPD. The difference between these coefficients, $\Delta\alpha_{\text{male-female}} = \alpha_{\text{male}} - \alpha_{\text{female}}$, represents the difference of the differential expression in males from that in females and allows us to quantify sexually dimorphic expression patterns in COPD relative to control tissue.

We can assess the statistical significance of the difference ($\Delta\alpha$) using Cohen's d, defined as the difference between two means divided by a pooled standard deviation: $d = \Delta\alpha / S_{\text{pooled}}$ [31,32]. As Cohen's d follows the standard normal distribution under the null, the two-tailed p-value is given as: $p_{\text{male-female}} = 2(1 - \Phi(|d|)) = 2(1 - \Phi(|\Delta\alpha / S_{\text{pooled}}|))$ where Φ is the cumulative distribution function of the standard normal distribution. Benjamini-Hochberg FDR was used for multiple hypothesis testing adjustment. A plot of these differences ($\Delta\alpha$) compared to the FDR significance is shown (Figure 4.1B).

We used an FDR cutoff of 0.25 to identify 1,551 genes that are most likely to be over-expressed in presence of COPD in one sex but relatively under-expressed or unchanged in the other. Our choice of the liberal FDR threshold optimizes for sensitivity [43], while the tradeoff in false discovery rate is minimized by requiring replication in a complementary COPD-stratified analysis below. In addition, we performed a sensitivity analysis on the choice of the FDR threshold, as described in Appendix 4A.

While stratified analyses are often used to study sex differences [44-46], we reasoned that by combining the COPD sex-stratified analysis above with results from a sex-differential expression analysis performed using stratification by COPD-control status we could increase

our confidence in our findings (Figure 4.1A). This COPD-only and control-only stratified analysis also allows us to identify gene expression that is not sexually dimorphic in control tissue but appears sexually dimorphic in the setting of COPD. Therefore, we quantified sex differences in controls and COPD cases separately by fitting the same basic model:

$$\text{Gene Expression} \mid \text{control} = \beta_{0,\text{control}} + \beta_{1,\text{control}}\,\text{Sex} + \beta_{2,\text{control}}\,\text{Pkyr smoked} + \beta_{3,\text{control}}\,\text{Age} + \varepsilon_{\text{control}} \quad (3)$$

$$\text{Gene Expression} \mid \text{COPD} = \beta_{0,\text{COPD}} + \beta_{1,\text{COPD}}\,\text{Sex} + \beta_{2,\text{COPD}}\,\text{Pkyr smoked} + \beta_{3,\text{COPD}}\,\text{Age} + \varepsilon_{\text{COPD}} \quad (4)$$

Here, the coefficient $\beta_{1,\text{control}}$ represents the sex-specific differences in gene expression in control lung tissue, accounting for age and pack years of smoking, while the coefficient $\beta_{1,\text{COPD}}$ represents sex-specific differences in COPD lung tissue. For simplicity, we will refer to $\beta_{1,\text{control}}$ as β_{control} and $\beta_{1,\text{COPD}}$ as β_{COPD}. A positive coefficient means higher gene expression in males compared to females, and a negative coefficient suggests higher gene expression in females compared to males. As in the sex-stratified analysis, sex-differences specific to either COPD or controls can be detected by contrasting the coefficients: $\Delta\beta_{\text{COPD-control}} = \beta_{\text{COPD}} - \beta_{\text{control}}$, and the significance: $p_{\text{COPD-control}} = 2(1 - \Phi(|\Delta\beta / S_{\text{pooled}}|))$ (see volcano plot in Figure S4.2). Of the autosomal Ensembl genes analyzed, 1,656 genes had significant sexually dimorphic differences at Benjamini-Hochberg false discovery rate of 0.25.

Identification of Sexually dimorphic and COPD differential (SDCD) genes

We intersected the significant gene sets identified in the sex-stratified and COPD stratified analyses to define a set of "sexually dimorphic and COPD differential genes" (SDCD genes, listed in Supplemental Table S2A). This overlap identified 959 SDCD genes; some of

these genes have been previously associated with COPD. Supplemental Table S3 and Figure S4.3 list a few notable examples including TIMP4 (top left of Figure 4.1B), a tissue inhibitor of metalloproteinase, which plays roles in mitigating degradation of lung matrix [47]. Consistent with a previous observation that TIMP4 is overexpressed in COPD [48], we found TIMP4 to be highly expressed in female COPD subjects compared to female controls ($\alpha_{female} = 0.249$, $p=1.15\times10^{-5}$); however, this differential expression is absent for males ($\alpha_{male} = 0.0143$, p=0.710, Figure 4.1C).

Among of 959 SDCD genes we also found genes with prior sex-specific associations such as AQP7, which has demonstrated sex-specific association with diabetes [49]. PAQR3 (Figure S4.3C), is among the most sexually dimorphic genes in COPD (based on the magnitude of β_{COPD}); this gene encodes a multiple-pass membrane which interacts with progesterone and has been identified in bronchoalveolar lavage fluid from ex-smokers [50]. At the individual gene level, this set of 959 SDCD genes represents a promising collection of genes relevant to both sexual dimorphism and COPD biology and forms the basis for our further analyses.

Validation of SDCD genes using independent datasets

To validate the SDCD gene set, we repeated our stratified gene expression analysis using independent gene expression datasets from two COPD case-control studies. From the Gene Expression Omnibus (GEO) database we selected datasets from Bhattacharya et al. (GEO#: GSE8581) [33] and Steiling et al. (GEO#: GSE37147) [28] based on the completeness of the clinical metadata and the sample sizes. In total, 411 SDCD genes were identified as differentially expressed in at least one independent dataset, and 78 were observed in both. A detailed

discussion of the datasets is provided in Appendix 4A, and the lists of SDCD genes are provided in the Supplemental Table S4.

Functional enrichment analysis of SDCD genes shows sex-specific expression of pathways

We performed functional enrichment analysis on the SDCD genes. We were motivated by the fact that significant enrichment for biological processes known to be important for COPD would add support for SDCD genes while additional processes may be relevant for sexually dimorphic features of the disease. To determine the enrichment of SDCD genes in Gene Ontology (GO) terms, we used the conditional hypergeometric test (Bioconductor Package GOstats [51] which accounts for the hierarchical structure of GO), and identified 181 terms with a significant over-representation of annotations from SDCD genes (defined as a p-value significance less than 0.05 and limited to terms containing three or more genes). Among the top GO terms is "anatomical structure morphogenesis" (GO:0009653), which may support the suggestion that sex differences in COPD are influenced by lung size and anatomy. Also key COPD processes such as "response to wounding" (GO:0009611) and "acute inflammatory response" (GO:0002526) are represented. A full list of these enriched GO terms is provided in Supplemental Tables S5A. We note that alternative approaches: pre-ranked Gene Set Enrichment Analysis (GSEA) [52] and DAVID [53,54], gave similar results and are included in Supplemental Tables S5B-D.

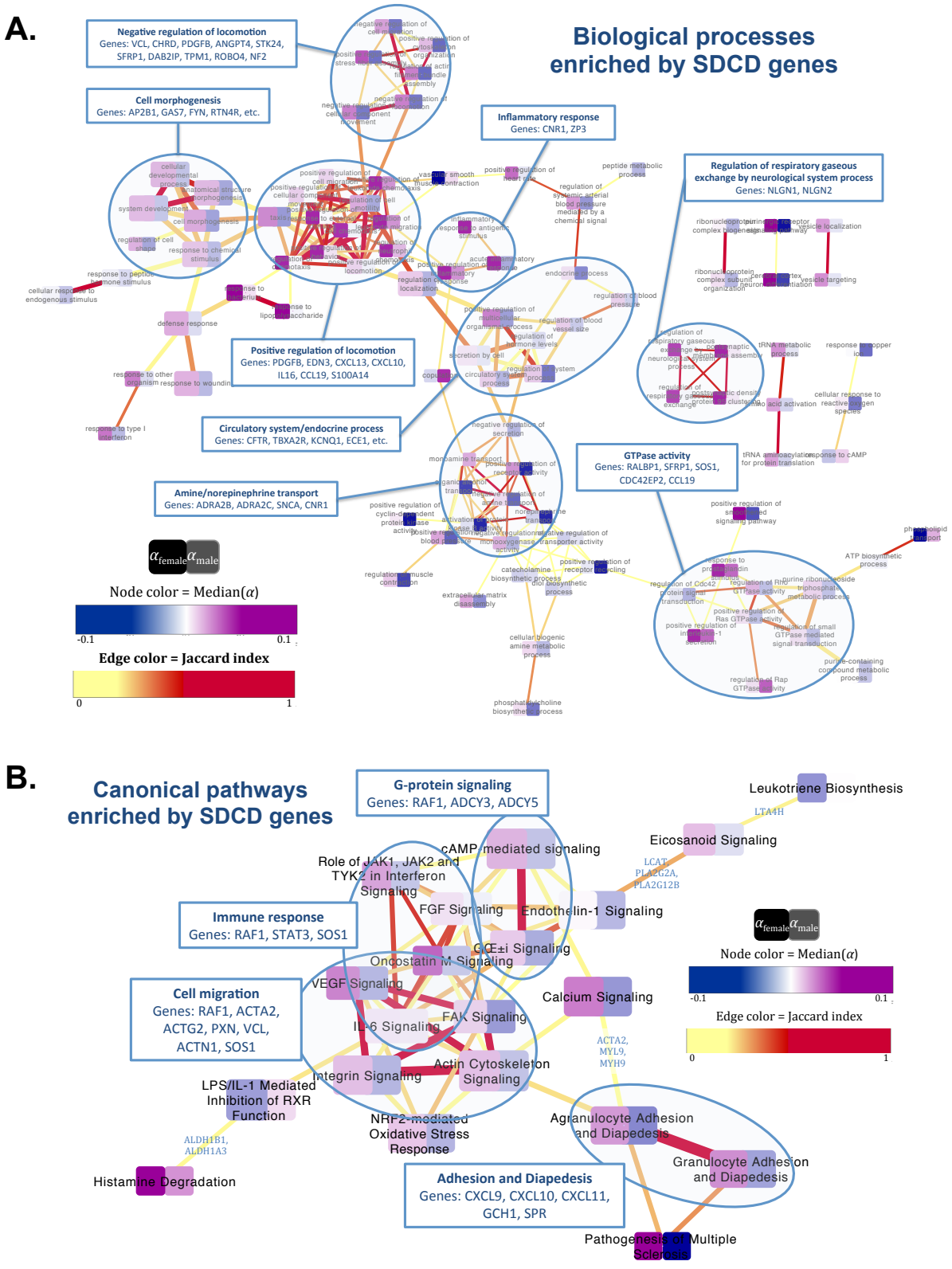


Figure 4.2: Network representation of Gene Ontology enrichment of SDCD genes. Nodes represent functions or pathways, colored by median sex-specific differential expression coefficients. Node size represents the total number of genes annotated to the functions.

(continued)
Edges represent the number of shared SDCD genes between pairs of functions or pathways, colored by the associated Jaccard index, which measures similarity between gene sets. The size of the edges represents the number of shared genes. The genes listed on each cluster are SDCD genes annotated in the majority of the terms in the cluster. (A) Through GLay "fast-greedy" network clustering [36], enriched biological processes are organized in groups with coherent themes such as cell morphogenesis, negative and positive regulation of locomotion, and GTPase activity. Some of these functions appear to be sex specific in their expression patterns. For example amine/norepinephrine transport is down-regulated primarily in male (blue nodes on the right) while positive regulation of locomotion is up-regulated primarily in female (pink nodes on the left). (B) Canonical pathways are also organized in modules. Each module has a coherent theme for example: VEGF signaling, actin cytoskeleton signaling, FAK signaling, and integrin signaling pathways are all involved in cell migration, while FGF signaling, IL6 signaling, oncostatin M signaling, and interferon signaling pathways are involved in immune responses. Two pathways, Histamine Degradation and Phenylalanine Degradation IV, do not share genes with other pathways and are not shown here.

Functional network organization

To visualize the functional enrichment results, we constructed a network representation of the significantly enriched GO terms, with edges defined based on shared SDCD genes between two functions (Figure 4.2A). We removed edges connecting functions to their subcategories to avoid simply recapitulating the GO hierarchy. We used the Jaccard index [55], which measures similarity between two gene sets as a proportion of the total shared genes, to define connectivity and only considered edges with a Jaccard index greater than or equal to 0.2 [56], suggesting the gene sets share a substantial number of genes.

To highlight trends of sex-specific differential expression of each function, we colored nodes in the network based on the median differential expression coefficients ($median(\alpha_{male})$ and $median(\alpha_{female})$) from male-only and female-only analyses. We used a fast-greedy clustering algorithm [36] to organize the network into several distinct groups of related biological processes, as shown in Figure S4.4 and highlighted in Figure 4.2A. For example, three terms

related to inflammatory response, a major contributing factor to COPD [57], form a small cluster in the middle of the network (Figure 4.2A). Chemotaxis and cell localization, other key processes in COPD, appear in two distinct clusters based on whether they are either positively or negatively regulated. Terms related to ATP biosynthesis process and GTPase activity are also represented in the network. An increase in extracellular ATP has been suggested to play a role in emphysema and COPD [58], potentially through altering chemotaxis and activation of inflammatory cells.

The SDCD genes associated with the same function are often differentially regulated in a sex-specific manner (node colors in Figure 4.2). For example, we find that genes involved in amine and norepinephrine transport are strongly down-regulated in COPD compared to control in males but are not significantly differentially expressed in females (lower center of Figure 4.2A). Genes involved in inflammatory response are strongly differentially expressed in females but not in males (upper center of Figure 4.2A). Genes involved in *positive* regulation of locomotion are over-expressed in males, while genes involved in *negative* regulation of locomotion are over-expressed in females (top of Figure 4.2A). As can be seen by the absence of edge connections between the two clusters, positive and negative regulations of locomotion involve distinct set of genes, and thus the agreement in the direction of differential expression between the two groups (positive regulation in male and negative regulation in female) suggests robustness of the sexually dimorphic signal.

Canonical Pathways

We performed a similar enrichment analysis on canonical pathways defined in the Ingenuity Pathway Knowledge Base and, using IPA (Ingenuity Systems, www.ingenuity.com) identified twenty-two pathways with significant associations with SDCD genes ($p<0.05$ with at least three SDCD genes, see Supplemental Table S5E), including the NRF2, VEGF, and IL6 pathways, all of which have been implicated in various aspects of COPD. After representing the pathways as a network (Figure 4.2B) in the same manner as the GO network, we found VEGF signaling, integrin signaling, actin cytoskeleton signaling, and FAK signaling pathways to be highly connected, sharing actin (ACTA2, ACTG2, ACTN1), paxillin (PXN), vinculin (VCL), SOS1, and RAF1. These pathways also have many connections to NRF2-mediated oxidative stress response through RAF1, ACTA2, and ACTG2. VEGF signaling is a crucial response to lung injury and VEGF deficiency has been associated with emphysema [59]. Murine models have implicated NRF2 in lung detoxification pathways for cigarette smoke and altered NRF2 activity has been associated with emphysema [60]. Another module comprises several immune response pathways including interferon signaling (JAK-STAT), IL-6 signaling, FGF signaling, and oncostatin M signaling pathways. Of note, both VEGF and IL-6 have demonstrated sexual dimorphism as plasma biomarkers in men and women with COPD and represent highly plausible sexually dimorphic COPD pathways [61].

Molecular mechanisms of sexual dimorphism

Sexually dimorphic gene expression may arise as a result of sex specific gene regulation. While regulation by sex hormones is a major mechanism of sex differences, genetic and epigenetic factors may also contribute. We used public databases of hormone responsive

elements, as well as genotypes and DNA methylation data from the LGRC to assess possible mechanisms of sex-specific regulation of the 959 SDCD genes. As described in more detail below, we found that hormones and DNA methylation likely play roles in sex-specific regulation of SDCD gene expression while genetic variation provides limited evidence.

The SDCD genes are enriched for hormone response elements

Hormones, especially sex hormones, contribute to sex differences in lung development and physiology [5,11,24,25]. Hormones regulate gene expression in part by activating their respective receptors, which then act as transcription factors, binding to specific hormone response elements (HREs) to alter gene transcription. Thus, if hormones are involved in gene regulation of SDCD genes, SDCD genes may harbor significantly higher number of hormone response elements than expected by chance.

Nuclear Receptor	Databases	SDCD genes targeted*	Target enrichment p-value
Estrogen (ESR1/2)	Cistrome [62,63], ERE DB [64], JASPAR [65]	284	0.0290
Androgen (AR)	Cistrome [66,67], ARGDB [68], JASPAR [65]	174	0.0165
Progesterone (PGR)	Cistrome [69]	47	0.0076
PPAR-gamma	Cistrome [70]	70	0.0136
Vitamin D (VDR)	Cistrome [71]	42	0.0391

Table 4.2: SDCD genes are enriched for regulation by hormone receptors. * The number of SDCD genes targeted by the nuclear receptor in at least one database. See Appendix 4A for details.

A number of studies have cataloged hormone response elements (see Appendix 4A for description). We used Fisher's exact test to assess significance of the number of SDCD genes harboring HREs curated in these databases and found there were more SDCD genes with HREs

than would be expected by chance (see Table 2). Because there are some content differences between HRE databases, we combined regulatory target gene lists from a number of databases. Here, we used the Cistrome database [62,63,72], ERE DB [64], and the JASPAR motif scan [65], and found an enrichment of SDCD genes regulated by estrogen receptor (ESR1/2, 284 genes, p=0.0290). Enrichment was also observed for the progesterone receptor with an elevated number of SDCD genes regulated by progesterone receptor in the Cistrome database [69] (47 genes, p = 0.0076).

Both progesterone and estrogen may influence COPD susceptibility through up-regulation of cytochrome P450 (CYP) enzymes [73,74]. Binding of estradiol to sex-hormone binding globulin (SHBG) triggers cAMP-dependent signaling, up-regulating adenyl cyclase (ADCY) and other downstream signaling molecules [75]. We observed that not only are cAMP-mediated signaling pathways enriched for SDCD genes (Supplemental Table S5E), but we also found adenyl cyclase 3 (ADCY3) and adenyl cyclase 5 (ADCY5) to be sexually dimorphic and differentially expressed in COPD.

We also observed significant results for androgen. Androgen receptor binding sites are overrepresented in SDCD genes (174 genes, p=0.0165) using three databases: Cistrome [66,67], ARGDB [68] and JASPAR motif scan targets [65]. One specific example is for SOD3 (extracellular superoxide dismutase 3); evidence for androgen regulation is observed for SOD3 in all three databases—importantly, SOD3 has been implicated in COPD and emphysema [76].

Finally, two nuclear receptors known to be important in COPD PPAR-gamma [21] and vitamin D receptor (VDR) [22,23] are also known to target SDCD genes. In the Cistrome

database [70,71], 42 and 70 SDCD genes are targeted by PPAR-gamma and VDR, respectively (p = 0.0391 and 0.0136, respectively). PPAR-gamma may function in a sex-specific manner [21] and has been proposed as a therapeutic target for COPD [20], as activation of PPAR-gamma attenuates inflammation [77]. Recent studies have found association between vitamin D deficiency and lung function [18], association between genetic variability in vitamin D binding protein and COPD severity [19] and emphysema [78]; reports of sex-specific effects of vitamin D in COPD have been limited.

Sexually dimorphic methylation

Variable DNA methylation has been associated with cigarette smoke exposure [79-82], and COPD [83], and is sex-specific in many tissues and cell types [84-88]. Since methylation of the regulatory sequences can alter gene transcription, sex-specific methylation may contribute to sexually dimorphic expression of SDCD genes.

We assessed sexual dimorphism of methylated sites within promoter regions of SDCD genes. Methylation data was available on 178 LGRC samples (134 COPD, 44 controls; 92 male, 86 female) on the Comprehensive High-throughput Arrays for Relative Methylation (CHARM) platform [37]. We normalized the data using R package CHARM, corrected for potential batch effects using surrogate variable analysis (SVA), and summarized as "variably methylated regions" (VMRs), as previously described [38,39] (see Methods). Of the 959 SDCD genes, 395 have VMRs within the promoter regions (defined as 10kb window around transcription start sites).

We detected sexually dimorphic and COPD-differential methylation patterns by stratified regression analysis, similar to the approach for the gene expression. We stratified the methylation profiles by COPD status, fitted COPD-only and control-only regression models, contrasted the regression coefficients through Cohen's *d*, and corrected the p-values using Benjamini-Hochberg FDR. At the 0.05 FDR cutoff, 387 VMRs, within promoter regions of 235 SDCD genes, showed significant sexually dimorphic and COPD-differential methylation (termed "SDCD VMRs"; Supplemental Table S6A).

One of the most significant SDCD VMRs is 6.5kb from calpain 12 (CAPN12, adjusted p-value = 0.002). Although there is little published data regarding a role for calpain 12 in the lung, calpain 12 has been implicated in apoptosis pathways and is a member of the general calpain superfamily. Calpains have also been described as mediating cigarette smoke induced angiogenic response [89]. In our data the CAPN12 VMR is hypermethylated in male controls compared to the other groups (Figure 4.3A). Moreover, the high level of methylation of the CAPN12 region in male controls corresponds with their low expression of CAPN12 mRNA (Figure 4.3B).

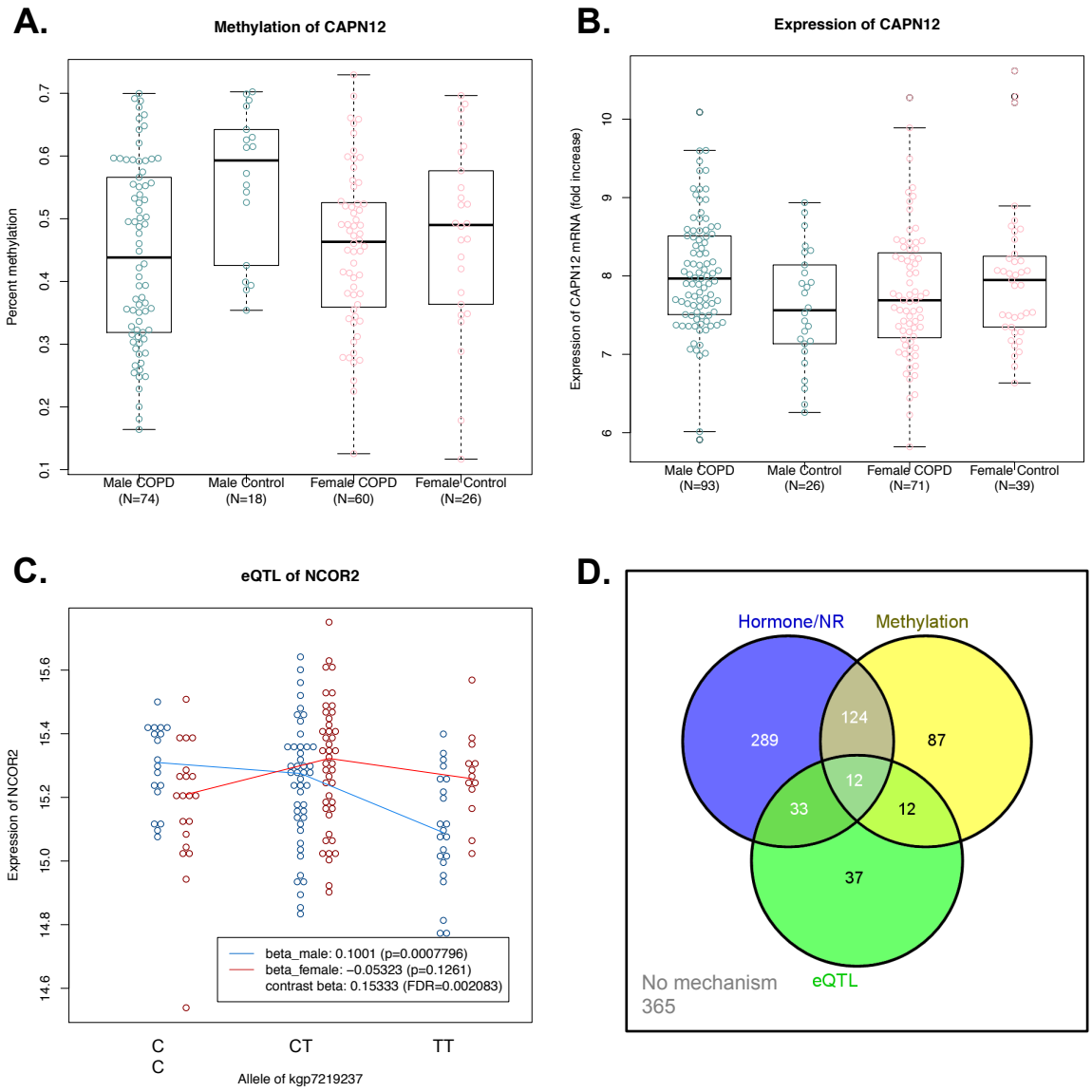


Figure 4.3: Regulatory mechanisms of SDCD gene expression. (A) Calpain 12 harbors a variably methylated region (VMR) that demonstrates sexually dimorphic methylation in controls. In particular, male control samples tend to be more highly methylated and male COPD cases relatively hypomethylated compared to male controls; females are equally methylated at a lower level. (B) Stratified boxplot of CAPN12 mRNA expression shows lowest expression in male controls, which is in agreement with their higher methylation level. (C) An example of a sexually dimorphic genetic regulation of gene expression (eQTL). Points represent gene expression of NCOR2 in male (blue) and female (red) in subjects with homozygous dominant (CC), heterozygous (CT), and homozygous recessive (TT). Lines connect median expression in each subgroup to demonstrate the trend. In males, expression of NCOR2 decreases as the number of minor allele increases, but there is no trend in females. (D) Combining all three lines of evidence for molecular mechanisms SDCD gene regulation, we provided hypotheses for 594 SDCD genes. Almost half of the SDCD genes (458) are targeted by sex hormone receptors, vitamin D receptor, or PPAR-gamma.

The SDCD genes with sexually dimorphic methylation are also functionally distinct. We performed functional enrichment analysis [53,54] of the 235 SDCD VMR genes relative to the background of all SDCD genes and found overrepresentation of genes associated with nucleotide binding (GO:0000166, 38 genes, adjusted p = 0.018), especially ATP binding (GO:0005524, 32 genes, adjusted p = 0.016). These include several ATPases, kinases, and ABCF2. ABDF2 is particularly interesting as it is a member of ATP-binding cassette family which includes genes for surfactant production and the cystic fibrosis transmembrane conductance regulator (CFTR) [90]. The full list of functional enrichment results is provided (Supplemental Table S6B).

Genetic markers show evidence for sexually dimorphic regulation of gene expression

Genetic variants can also affect gene expression by modifying transcription factor binding sites and so can alter response to co-factors, including sex-specific factors such as androgen and estrogen. Expression quantitative trait loci (eQTL) analysis can be used to identify potential regulatory variants. We performed sex-stratified *cis*-eQTL analysis, in which gene expression is modeled as a linear function of the additive effect of a SNP, adjusting for age and pack years of smoking (see Methods). As in the SDCD gene expression analysis, we contrasted the coefficients of male-only and female-only eQTL models.

A screen of 28,301 SNPs within 100kb upstream and 10kb downstream of the transcription start sites (TSS) of SDCD genes identified 302 variant positions associated with 94 genes that were statistically significant, sexually dimorphic eQTLs (Supplemental Table S6C). Here we used an FDR cutoff of 0.05 and required that the SNPs had more than five samples

with homozygous recessive alleles in each sex, to mitigate the influence of outliers. Among the top sexually dimorphic eQTL are MMP15 (adjusted p=0.002), an important matrix protease whose genetic variation is associated with lung function [91,92]; NCOR2 (adjusted p=0.002), which interacts with HDAC3, implicated in inflammatory response to cigarette smoke [93], PXN (adjusted p=0.003), a key signaling protein in VEGF pathway [94]; and TIMP4 (adjusted p=0.011), an inhibitor of MMP proteins found to be overexpressed in COPD [48].

Figure 4.3C shows the relationship between the number of minor alleles of SNP kgp7219237 and gene expression of NCOR2, a gene that has been implicated as required for terminal lung development [95], potentially a point when sexual dimorphism in the lung may begin. NCOR2, nuclear receptor corepressor 2 or silencing mediator of retinoic acid and thyroid hormone receptor (SMRT), also functions by recruiting histone deacetylases (HDACs), which may suppress inflammation in COPD [96,97]. Here, the expression of NCOR2 mRNA decreases in males as a function of the number of minor allele, but the NCOR2 expression in females does not appear to change significantly (Figure 4.3C).

Sensitivity Analyses

Although we made an effort to adjudicate the clinical classification of the LGRC study cohort based on available data, some samples may be incorrectly classified. In particular, a subset of control samples have atypically low diffusion capacity (DLCO; N=24), and about 14% (N=32) of the LGRC samples self-reported never having smoked cigarettes. Therefore we repeated our sexually dimorphic differential expression analysis to assess the sensitivity of our results to the presence of these samples. We found that the SDCD genes and their functional

enrichments are robust to the inclusion or exclusion of these samples, with majority of the SDCD genes replicated in both sensitivity analyses (see Appendix 4A and Supplemental Tables S7A-D).

We also tested the sensitivity of our results to FDR threshold and effect size. To assess our choice of FDR threshold, we reran the stratified differential expression analyses varying the FDR threshold between 0.01 and 1 and found the significance of the overlap between sex-stratified and COPD-stratified analyses is optimal at the FDR threshold between 0.1 and 0.25, thus supporting our choice of 0.25 (see Appendix 4A). To evaluate the relevance of effect size in the definition of SDCD genes, we selected the top 5% based on the magnitude of differential expression $|\Delta\alpha_{\text{male-female}}|$ and $|\Delta\beta_{\text{COPD-control}}|$ and found their functional enrichments to be consistent with that of the full SDCD gene set (see Appendix 4A and Supplemental Tables S7E-F). This consistency suggests that the full set of SDCD genes is at least as functionally relevant as the set with large effect size. Appendix 4A contains a detailed discussion of the sensitivity analyses, and the lists of genes and their functional enrichments can be found in Supplemental Table S7. Taken together sensitivity analyses affirm that SDCD genes and their functional enrichments are robust against clinical misclassification and model parameter choices.

Discussion

Sex and gender disparities in COPD have been recognized for several decades, highlighting the importance of sex- and gender-based research that may yield insights into sex-specific treatments for COPD. Evidence suggests that females may be more susceptible to COPD, yet little is known about the molecular mechanisms underlying the sexual

dimorphism—information with potential translational relevance for sex-specific therapeutics. Our analysis represents the first systematic genome-wide integrative genomics analysis of sexual dimorphism in COPD. Using gene expression profiles from COPD and control samples from the Lung Genomics Research Consortium (LGRC), we identified 959 autosomal genes with evidence for sexually dimorphic and COPD differential expression (SDCD genes); 150 of these genes are not dimorphic at baseline between men and women and demonstrate dimorphism in the presence of COPD, perhaps highlighting the most relevant sex-specific disease targets.

The 959 genes are involved in key functions including cell morphogenesis, inflammatory response, and regulation of chemotaxis, as well as functions not typically associated with COPD such as GTPase activity, amine transport, and endocrine processes. By using publicly available hormone response element databases and methylation and genotyping data from the LGRC, we found potential regulatory mechanisms involving hormonal regulation and DNA methylation that that may be associated with the sexually dimorphic gene expression patterns we observed.

Of the 959 SDCD genes, a few are particularly interesting. Lung tissue remodeling and faulty repair in response to cigarette smoke are important in COPD pathogenesis [89] as impaired tissue repair process may lead to permanent structural remodeling of the lung, a key characteristic of COPD. Calpains, a family of calcium-dependent intracellular proteases, have been implicated in angiogenic response to cigarette smoke [89]. Here we found CAPN12 to exhibit strong sexual dimorphism in both gene expression and methylation. The pattern of differential expression of CAPN12 inversely associated with methylation (Figure 4.3A-B). Since cigarette smoke-induced inhibition of calpain could lead to an impaired tissue remodeling,

down-regulation of CAPN12 in females may suggest an inadequate response to cigarette smoke among females, in line with the epidemiologic observations. Since there is not much literature for a role for this calpain in the lung more research is needs to define mechanistic relevance of our observation.

Additional interesting candidates include the matrix metalloproteinases (MMP3, MMP15, MMP28) and their inhibitor TIMP4 (Figure S4.3). MMPs, extracellular matrix enzymes, and their inhibitor TIMPs play roles in COPD disease progression via degradation of elastin and aberrant lung tissue remodeling [47]. MMP3 (stromolysin; Figure S4.3B) and TIMP4 (Figure 4.1C) are among the most statistically significant SDCD genes in our expression analysis and have been investigated in association with airway inflammation and remodeling [47] as well as COPD development and progression [98,99]. While a previous report found overexpression of MMP3 and TIMP4 in COPD [48], our analysis suggests that this might be driven by the direction and magnitude of effect in females (Figures 1C and S3B). Variants in the MMP15 gene have been associated with lung function in two genome-wide association studies [91,92], and MMP15 was highlighted in our eQTL analysis (Figure 4.3C). MMP28 (epilysin) is overexpressed in response to injury [100] and has been implicated in airway epithelial cell survival [101]. In our analysis, MMP28 is one of the genes whose expressions are sexually dimorphic only in COPD, not among controls (Figure S4.3F). The fact that three members of MMP family and their inhibitor TIMP4 exhibit sexually dimorphic differential expression presents a strong case for sex-specificity of the roles of MMP and TIMP in COPD. Inhibition of MMPs has been proposed

as a potential therapeutic for COPD [102]; thus the finding of sexual dimorphism in MMPs supports the relevance of our results toward considerations of sex-specific therapeutics.

A total of 150 of the 959 SDCD were sexually dimorphic in the presence of COPD only (Supplemental Table S2B). Genes meeting these criteria include AQP7, which has demonstrated sex-specific association with diabetes [49], and TRIM40, a Class I MHC gene that has been associated with multiple sclerosis [103]. PAQR3 (Figure S4.3C), is also among the most sexually dimorphic genes in COPD (based on the magnitude of β_{COPD}); this gene encodes a multiple-pass membrane protein that interacts with progesterone and has been identified in bronchoalveolar lavage fluid from ex-smokers [50]. One caveat about this set of genes is that their expression may not appear significantly sexually dimorphic in control tissues because of the smaller sample size of controls compared to COPD (75 vs 179).

The pathways enriched in SDCD genes include NRF2, IL6 and VEGF signaling. Vascular endothelial growth factor (VEGF) signaling is of particular interest as VEGF is a key mediator of angiogenesis, with associated effects on lung development and pulmonary physiology [104,105] and has been observed to demonstrate sex differences as a plasma biomarker in men and women with COPD [61]. In particular, VEGF-activated cell migration pathway (left most path in Figure 4.4A-B) appears to be highly sexually dimorphic, with four of the five components (actin, alpha actin, paxillin, vinculin) of the signaling protein complex classified as SDCD genes.

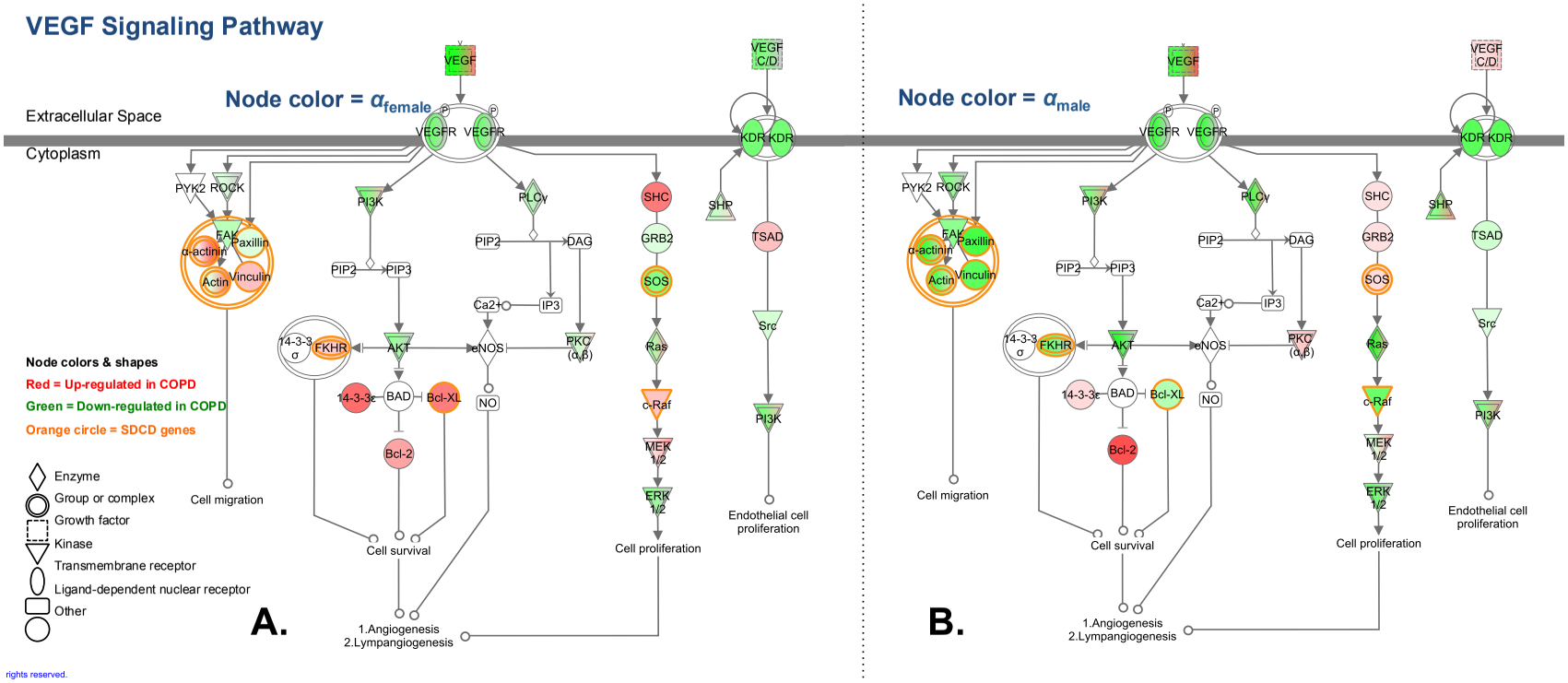


Figure 4.4: VEGF signaling pathway represented by Ingenuity Pathway Analysis (IPA) tool. Each molecule in the pathways is painted by the coefficient of differential gene expression in (A) female-only analysis (α_{female}) and (B) male-only analysis (α_{male}). Green = down-regulated in COPD, Red = up-regulated in COPD. The genes marked by orange edges are SDCD genes. The complex on the left contains four SDCD genes: actin, alpha actin, paxillin, and vinculin. This complex interacts with integrin and triggers cell migration.

VEGF signaling, integrin signaling, and actin signaling pathways share a number of genes (Figure 4.2B). A closer look reveals that the actin-alpha actin-paxillin-vinculin complex is a common feature among the three pathways. In the VEGF signaling pathway (Figure 4.4) and actin cytoskeleton signaling pathway (Figure S4.6A-B), the complex triggers downstream focal adhesion assembly and cell migration, while in the integrin signaling pathway (Figure S4.6C-D), the complex is an important component of integrin receptor aggregation, which stimulates signal transduction cascades. It is possible that the VEGF signaling and integrin signaling pathways may function together, and further *in vitro* research is needed to investigate this potential dimorphic signaling cascade.

Integrating the mechanistic and functional information gives a more complete view of the regulation of VEGF signaling pathway. Of the eleven SDCD genes involved in VEGF signaling pathway, as defined in Ingenuity Pathway Knowledge Base, two contained estrogen response elements (BCL2L1 and ACTN1). The connection between estrogen and VEGF signaling pathway has been observed during lung development [106], in thyroid tumors [107], and breast cancer [108], in which an estrogen-mediated angiogenesis functions through the VEGF pathway [108]; to date, there is minimal research specifically on estrogen-VEFG interactions and COPD.

A few limitations exist in our study. First many, but not all, of the subjects had lung cancer. Although the lung tissues were taken from areas distal to the tumors, the potential field effect associated with lung cancer could impact our findings. Second, since the tissue samples are whole lung homogenates, it is not possible to resolve the impact of cell type composition on

the results. Due to potential variation in cell type composition across control, COPD, male, and female lungs, the differential expression signatures may be driven by cellular heterogeneity and is the focus of ongoing research.

Although we are able to validate sexually dimorphic expression of 78 genes in two independent datasets, a number of SDCD genes are not replicated. This is possibly due to a few key characteristic differences among the datasets. For instance, Steiling and colleagues collected samples from airway epithelial brushings, a different tissue type than the LGRC. If cell types other than the airway epithelium are major drivers of COPD sexual dimorphism, one may not expect the results to replicate well in Steiling's dataset. Since about a quarter of SDCD genes were replicable in the Steiling's dataset, we believe airway epithelial cells partly contribute to the sexually dimorphic signal, but other cell types are likely important as well.

Our study begins to address the complexity of molecular contributions to sex differences in COPD by systematically investigating sexually dimorphic differential expression signatures and integrating these findings with regulatory information. This analysis leverages the Lung Genomics Research Consortium rich data, and the list of 959 SDCD genes can provide as an entry point for future functional studies in COPD to further elucidate sex-specific biology underlying this devastating lung disease. Our study highlights sexually dimorphic features of some pathways already implicated in COPD, including VEGF, IL6 and NRF2 signaling pathways, and uncovers other pathways not traditionally associated with COPD such as GTPase activity, amine transport, and endocrine processes. While mechanisms such as methylation and hormone regulation shed light on the complex nature of COPD susceptibility

and severity, integrative genomics approaches may offer hope for elucidation of sex-specific therapeutic targets. Our results stress the importance of considering sex as a key factor in the development of diagnostic, therapeutic, and preventative strategies.

Chapter 4 Bibliography

1. Hoyert DL, Arias E, Smith BL, Murphy SL, Kochanek KD (2001) Deaths: final data for 1999. Natl Vital Stat Rep 49: 1-113.

2. Balkissoon R, Lommatzsch S, Carolan B, Make B (2011) Chronic obstructive pulmonary disease: a concise review. Med Clin North Am 95: 1125-1141.

3. Ford ES, Croft JB, Mannino DM, Wheaton AG, Zhang X, et al. (2013) COPD surveillance--United States, 1999-2011. Chest 144: 284-305.

4. Mannino DM, Homa DM, Akinbami LJ, Ford ES, Redd SC (2002) Chronic obstructive pulmonary disease surveillance--United States, 1971-2000. Respir Care 47: 1184-1199.

5. Sin DD, Cohen SB, Day A, Coxson H, Pare PD (2007) Understanding the biological differences in susceptibility to chronic obstructive pulmonary disease between men and women. Proc Am Thorac Soc 4: 671-674.

6. Han MK, Postma D, Mannino DM, Giardino ND, Buist S, et al. (2007) Gender and chronic obstructive pulmonary disease: why it matters. Am J Respir Crit Care Med 176: 1179-1184.

7. Hemsing N, Greaves L (2009) Women, environments and chronic disease: shifting the gaze from individual level to structural factors. Environ Health Insights 2: 127-135.

8. Silverman EK, Weiss ST, Drazen JM, Chapman HA, Carey V, et al. (2000) Gender-related differences in severe, early-onset chronic obstructive pulmonary disease. Am J Respir Crit Care Med 162: 2152-2158.

9. Martinez FJ, Curtis JL, Sciurba F, Mumford J, Giardino ND, et al. (2007) Sex differences in severe pulmonary emphysema. Am J Respir Crit Care Med 176: 243-252.

10. Burrows B, Bloom JW, Traver GA, Cline MG (1987) The course and prognosis of different forms of chronic airways obstruction in a sample from the general population. N Engl J Med 317: 1309-1314.

11. Becklake MR, Kauffmann F (1999) Gender differences in airway behaviour over the human life span. Thorax 54: 1119-1138.

12. Gold DR, Wang X, Wypij D, Speizer FE, Ware JH, et al. (1996) Effects of cigarette smoking on lung function in adolescent boys and girls. N Engl J Med 335: 931-937.

13. Gan WQ, Man SF, Postma DS, Camp P, Sin DD (2006) Female smokers beyond the perimenopausal period are at increased risk of chronic obstructive pulmonary disease: a systematic review and meta-analysis. Respir Res 7: 52.

14. Akbas T, Karakurt S, Unluguzel G, Celikel T, Akalin S (2010) The endocrinologic changes in critically ill chronic obstructive pulmonary disease patients. COPD 7: 240-247.

15. Svartberg J (2010) Androgens and chronic obstructive pulmonary disease. Curr Opin Endocrinol Diabetes Obes 17: 257-261.

16. Chavoshan B, Fournier M, Lewis MI, Porszasz J, Storer TW, et al. (2012) Testosterone and resistance training effects on muscle nitric oxide synthase isoforms in COPD men. Respir Med 106: 269-275.

17. Atlantis E, Fahey P, Cochrane B, Wittert G, Smith S (2013) Endogenous testosterone level and testosterone supplementation therapy in chronic obstructive pulmonary disease (COPD): a systematic review and meta-analysis. BMJ Open 3: e003127.

18. Black PN, Scragg R (2005) Relationship between serum 25-hydroxyvitamin d and pulmonary function in the third national health and nutrition examination survey. Chest 128: 3792-3798.

19. Janssens W, Bouillon R, Claes B, Carremans C, Lehouck A, et al. (2010) Vitamin D deficiency is highly prevalent in COPD and correlates with variants in the vitamin D-binding gene. Thorax 65: 215-220.

20. Belvisi MG, Hele DJ (2008) Peroxisome proliferator-activated receptors as novel targets in lung disease. Chest 134: 152-157.

21. Corbo RM, Pinto A, Scacchi R (2013) Gender-specific association between FSHR and PPARG common variants and human longevity. Rejuvenation Res 16: 21-27.

22. Dam TT, von Muhlen D, Barrett-Connor EL (2009) Sex-specific association of serum vitamin D levels with physical function in older adults. Osteoporos Int 20: 751-760.

23. Fereidan-Esfahani M, Ramagopalan SV, Etemadifar M, Sadri S, Abtahi SH (2013) Vitamin d: shining a light on clinical and sex specific effects in multiple sclerosis? Int J Prev Med 4: 499-500.

24. Mannino DM, Holguin F, Greves HM, Savage-Brown A, Stock AL, et al. (2004) Urinary cadmium levels predict lower lung function in current and former smokers: data from the Third National Health and Nutrition Examination Survey. Thorax 59: 194-198.

25. Ben-Zaken Cohen S, Pare PD, Man SF, Sin DD (2007) The growing burden of chronic obstructive pulmonary disease and lung cancer in women: examining sex differences in cigarette smoke metabolism. Am J Respir Crit Care Med 176: 113-120.

26. Chen ZH, Kim HP, Ryter SW, Choi AM (2008) Identifying targets for COPD treatment through gene expression analyses. Int J Chron Obstruct Pulmon Dis 3: 359-370.

27. Ning W, Li CJ, Kaminski N, Feghali-Bostwick CA, Alber SM, et al. (2004) Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease. Proc Natl Acad Sci U S A 101: 14895-14900.

28. Steiling K, van den Berge M, Hijazi K, Florido R, Campbell J, et al. (2013) A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. Am J Respir Crit Care Med 187: 933-942.

29. Wang IM, Stepaniants S, Boie Y, Mortimer JR, Kennedy B, et al. (2008) Gene expression profiling in patients with chronic obstructive pulmonary disease and lung cancer. Am J Respir Crit Care Med 177: 402-411.

30. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3: 1724-1735.

31. Cohen J (1988) Statistical power analysis for the behavioral sciences. Hillsdale, N.J.: L. Erlbaum Associates. xxi, 567 p. p.

32. Nunn CL, Lindenfors P, Pursall ER, Rolff J (2009) On sexual dimorphism in immune function. Philos Trans R Soc Lond B Biol Sci 364: 61-69.

33. Bhattacharya S, Srisuma S, Demeo DL, Shapiro SD, Bueno R, et al. (2009) Molecular biomarkers for quantitative and discrete COPD phenotypes. Am J Respir Cell Mol Biol 40: 359-367.

34. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.

35. Noack A (2009) Modularity clustering is force-directed layout. Phys Rev E Stat Nonlin Soft Matter Phys 79: 026102.

36. Su G, Kuchinsky A, Morris JH, States DJ, Meng F (2010) GLay: community structure analysis of biological networks. Bioinformatics 26: 3135-3137.

37. Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, et al. (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res 18: 780-790.

38. Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, et al. (2010) Personalized epigenomic signatures that are stable over time and covary with body mass index. Sci Transl Med 2: 49ra67.

39. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT (2012) Significance analysis and statistical dissection of variably methylated regions. Biostatistics 13: 166-178.

40. Wang S, Sun H, Ma J, Zang C, Wang C, et al. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protoc 8: 2502-2515.

41. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

42. Baker DA, Meadows LA, Wang J, Dow JA, Russell S (2007) Variable sexually dimorphic gene expression in laboratory strains of Drosophila melanogaster. BMC Genomics 8: 454.

43. Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375.

44. Browning BL, Annese V, Barclay ML, Bingham SA, Brand S, et al. (2008) Gender-stratified analysis of DLG5 R30Q in 4707 patients with Crohn disease and 4973 controls from 12 Caucasian cohorts. J Med Genet 45: 36-42.

45. Randall JC, Winkler TW, Kutalik Z, Berndt SI, Jackson AU, et al. (2013) Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. PLoS Genet 9: e1003500.

46. Taylor KC, Carty CL, Dumitrescu L, Buzkova P, Cole SA, et al. (2013) Investigation of gene-by-sex interactions for lipid traits in diverse populations from the population architecture using genomics and epidemiology study. BMC Genet 14: 33.

47. Lagente V, Manoury B, Nenan S, Le Quement C, Martin-Chouly C, et al. (2005) Role of matrix metalloproteinases in the development of airway inflammation and remodeling. Braz J Med Biol Res 38: 1521-1530.

48. Navratilova Z, Zatloukal J, Kriegova E, Kolek V, Petrek M (2012) Simultaneous up-regulation of matrix metalloproteinases 1, 2, 3, 7, 8, 9 and tissue inhibitors of metalloproteinases 1, 4 in serum of patients with chronic obstructive pulmonary disease. Respirology 17: 1006-1012.

49. Prudente S, Flex E, Morini E, Turchi F, Capponi D, et al. (2007) A functional variant of the adipocyte glycerol channel aquaporin 7 gene is associated with obesity and related metabolic abnormalities. Diabetes 56: 1468-1474.

50. Zhan X, Desiderio DM (2011) Nitroproteins Identified in Human Ex-smoker Bronchoalveolar Lavage Fluid. Aging Dis 2: 100-115.

51. Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. Bioinformatics 23: 257-258.

52. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545-15550.

53. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

54. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.

55. Jaccard P (1912) The distribution of the flora in the alpine zone. New Phytologist 11: 37-50.

56. Ye H, Liu Q, Wei J (2014) Construction of drug network based on side effects and its application for drug repositioning. PLoS One 9: e87864.

57. Cornwell WD, Kim V, Song C, Rogers TJ (2010) Pathogenesis of inflammation and repair in advanced COPD. Semin Respir Crit Care Med 31: 257-266.

58. Mortaz E, Folkerts G, Nijkamp FP, Henricks PA (2010) ATP and the pathogenesis of COPD. Eur J Pharmacol 638: 1-4.

59. Lee CG, Ma B, Takyar S, Ahangari F, Delacruz C, et al. (2011) Studies of vascular endothelial growth factor in asthma and chronic obstructive pulmonary disease. Proc Am Thorac Soc 8: 512-515.

60. Rangasamy T, Cho CY, Thimmulappa RK, Zhen L, Srisuma SS, et al. (2004) Genetic ablation of Nrf2 enhances susceptibility to cigarette smoke-induced emphysema in mice. J Clin Invest 114: 1248-1259.

61. de Torres JP, Casanova C, Pinto-Plata V, Varo N, Restituto P, et al. (2011) Gender differences in plasma biomarker levels in a cohort of COPD patients: a pilot study. PloS one 6: e16021.

62. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, et al. (2006) Genome-wide analysis of estrogen receptor binding sites. Nat Genet 38: 1289-1297.

63. Vivar OI, Zhao X, Saunier EF, Griffin C, Mayba OS, et al. (2010) Estrogen receptor beta binds to and regulates three distinct classes of target genes. J Biol Chem 285: 22059-22066.

64. Bourdeau V, Deschenes J, Metivier R, Nagai Y, Nguyen D, et al. (2004) Genome-wide identification of high-affinity estrogen response elements in human and mouse. Mol Endocrinol 18: 1411-1427.

65. Sandelin A, Wasserman WW (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol 338: 207-215.

66. Yu J, Yu J, Mani RS, Cao Q, Brenner CJ, et al. (2010) An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell 17: 443-454.

67. Wang Q, Li W, Zhang Y, Yuan X, Xu K, et al. (2009) Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. Cell 138: 245-256.

68. Jiang M, Ma Y, Chen C, Fu X, Yang S, et al. (2009) Androgen-responsive gene database: integrated knowledge on androgen-responsive genes. Mol Endocrinol 23: 1927-1933.

69. Tang Q, Chen Y, Meyer C, Geistlinger T, Lupien M, et al. (2011) A comprehensive view of nuclear receptor cancer cistromes. Cancer Res 71: 6940-6947.

70. Lefterova MI, Zhang Y, Steger DJ, Schupp M, Schug J, et al. (2008) PPARgamma and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale. Genes Dev 22: 2941-2952.

71. Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, et al. (2010) A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome Res 20: 1352-1360.

72. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, et al. (2011) Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol 12: R83.

73. Dimitropoulou C, Drakopanagiotakis F, Catravas JD (2007) Estrogen as a new therapeutic target for asthma and chronic obstructive pulmonary disease. Drug News Perspect 20: 241-252.

74. Tam A, Morrish D, Wadsworth S, Dorscheid D, Man SF, et al. (2011) The role of female hormones on lung function in chronic lung diseases. BMC Womens Health 11: 24.

75. Nakhla AM, Khan MS, Romas NP, Rosner W (1994) Estradiol causes the rapid accumulation of cAMP in human prostate. Proc Natl Acad Sci U S A 91: 5402-5405.

76. Sorheim IC, DeMeo DL, Washko G, Litonjua A, Sparrow D, et al. (2010) Polymorphisms in the superoxide dismutase-3 gene are associated with emphysema in COPD. COPD 7: 262-268.

77. Standiford TJ, Keshamouni VG, Reddy RC (2005) Peroxisome proliferator-activated receptor-{gamma} as a regulator of lung inflammation and repair. Proc Am Thorac Soc 2: 226-231.

78. Berg I, Hanson C, Sayles H, Romberger D, Nelson A, et al. (2013) Vitamin D, vitamin D binding protein, lung function and structure in COPD. Respir Med 107: 1578-1588.

79. Zhang FF, Cardarelli R, Carroll J, Fulda KG, Kaur M, et al. (2011) Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. Epigenetics 6: 623-629.

80. Ehrlich S, Walton E, Roffman JL, Weiss D, Puls I, et al. (2012) Smoking, but not malnutrition, influences promoter-specific DNA methylation of the proopiomelanocortin gene in patients with and without anorexia nervosa. Can J Psychiatry 57: 168-176.

81. Gilson E, Horard B (2012) Comprehensive DNA methylation profiling of human repetitive DNA elements using an MeDIP-on-RepArray assay. Methods Mol Biol 859: 267-291.

82. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, et al. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet 21: 3073-3082.

83. Qiu W, Baccarelli A, Carey VJ, Boutaoui N, Bacherman H, et al. (2012) Variable DNA methylation is associated with chronic obstructive pulmonary disease and lung function. Am J Respir Crit Care Med 185: 373-381.

84. Gabory A, Attig L, Junien C (2009) Sexual dimorphism in environmental epigenetic programming. Mol Cell Endocrinol 304: 8-18.

85. Vaissiere T, Hung RJ, Zaridze D, Moukeria A, Cuenin C, et al. (2009) Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant DNA methylation of specific genes and its association with gender and cancer risk factors. Cancer Res 69: 243-252.

86. El-Maarri O, Becker T, Junen J, Manzoor SS, Diaz-Lacava A, et al. (2007) Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. Hum Genet 122: 505-514.

87. Semaan SJ, Dhamija S, Kim J, Ku EC, Kauffman AS (2012) Assessment of epigenetic contributions to sexually-dimorphic Kiss1 expression in the anteroventral periventricular nucleus of mice. Endocrinology 153: 1875-1886.

88. Kurian JR, Olesen KM, Auger AP (2010) Sex differences in epigenetic regulation of the estrogen receptor-alpha promoter within the developing preoptic area. Endocrinology 151: 2297-2305.

89. Su Y, Cao W, Han Z, Block ER (2004) Cigarette smoke extract inhibits angiogenesis of pulmonary artery endothelial cells: the role of calpain. Am J Physiol Lung Cell Mol Physiol 287: L794-800.

90. van der Deen M, de Vries EG, Timens W, Scheper RJ, Timmer-Bosscha H, et al. (2005) ATP-binding cassette (ABC) transporters in normal and pathological lung. Respir Res 6: 59.

91. Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, et al. (2011) Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet 43: 1082-1090.

92. Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, et al. (2009) A genome-wide association study of pulmonary function measures in the Framingham Heart Study. PLoS Genet 5: e1000429.

93. Winkler AR, Nocka KN, Williams CM (2012) Smoke exposure of human macrophages reduces HDAC3 activity, resulting in enhanced inflammatory cytokine production. Pulm Pharmacol Ther 25: 286-292.

94. Abedi H, Zachary I (1997) Vascular endothelial growth factor stimulates tyrosine phosphorylation and recruitment to new focal adhesions of focal adhesion kinase and paxillin in endothelial cells. J Biol Chem 272: 15442-15451.

95. Pei L, Leblanc M, Barish G, Atkins A, Nofsinger R, et al. (2011) Thyroid hormone receptor repression is linked to type I pneumocyte-associated respiratory distress syndrome. Nat Med 17: 1466-1472.

96. Delcuve GP, Khan DH, Davie JR (2012) Roles of histone deacetylases in epigenetic regulation: emerging paradigms from studies with inhibitors. Clin Epigenetics 4: 5.

97. Barnes PJ (2006) Reduced histone deacetylase in COPD: clinical implications. Chest 129: 151-155.

98. Ohno I, Ohtani H, Nitta Y, Suzuki J, Hoshi H, et al. (1997) Eosinophils as a source of matrix metalloproteinase-9 in asthmatic airway inflammation. Am J Respir Cell Mol Biol 16: 212-219.

99. Korytina GF, Tselousova OS, Akhmadishina LZ, Victorova EV, Zagidullin Sh Z, et al. (2012) [Association of the MMP3, MMP9, ADAM33 and TIMP3 genes polymorphic markers with development and progression of chronic obstructive pulmonary disease]. Mol Biol (Mosk) 46: 487-499.

100. Lohi J, Wilson CL, Roby JD, Parks WC (2001) Epilysin, a novel human matrix metalloproteinase (MMP-28) expressed in testis and keratinocytes and in response to injury. J Biol Chem 276: 10134-10144.

101. Manicone AM, Harju-Baker S, Johnston LK, Chen AJ, Parks WC (2011) Epilysin (matrix metalloproteinase-28) contributes to airway epithelial cell survival. Respir Res 12: 144.

102. Daheshia M (2005) Therapeutic inhibition of matrix metalloproteinases for the treatment of chronic obstructive pulmonary disease (COPD). Curr Med Res Opin 21: 587-594.

103. Cree BA, Rioux JD, McCauley JL, Gourraud PA, Goyette P, et al. (2010) A major histocompatibility Class I locus contributes to multiple sclerosis susceptibility independently from HLA-DRB1*15:01. PLoS One 5: e11296.

104. Knox AJ, Stocks J, Sutcliffe A (2005) Angiogenesis and vascular endothelial growth factor in COPD. Thorax 60: 88-89.

105. Kanazawa H (2007) Role of vascular endothelial growth factor in the pathogenesis of chronic obstructive pulmonary disease. Med Sci Monit 13: RA189-195.

106. Gortner L, Shen J, Tutdibi E (2013) Sexual dimorphism of neonatal lung development. Klin Padiatr 225: 64-69.

107. Kamat A, Rajoria S, George A, Suriano R, Shanmugam A, et al. (2011) Estrogen-mediated angiogenesis in thyroid tumor microenvironment is mediated through VEGF signaling pathways. Arch Otolaryngol Head Neck Surg 137: 1146-1153.

108. Applanat MP, Buteau-Lozano H, Herve MA, Corpet A (2008) Vascular endothelial growth factor is a target gene for estrogen receptor and contributes to breast cancer progression. Adv Exp Med Biol 617: 437-444.

Chapter 5: Bicluster-based error rate control for eQTL analysis—BBER

Abstract

Background: Expression quantitative trait loci (eQTL) analysis combines genotype and mRNA expression data to identify genetic variants that have a gene-regulatory association. Because millions of genetic variants and tens of thousands of measured gene expression profiles are compared, adjusting for multiple testing is essential to assure reliable results. Widely-used approaches such as the Benjamini-Hochberg procedure largely assume independence between tests and ignore intrinsic and strong correlations among genetic and expression markers, resulting in suboptimal sensitivity.

Results: BBER extends Benjamini-Hochberg procedure by grouping correlated tests based on genetic linkage disequilibrium and co-expression pattern, thus reducing the apparent number of independent tests and mitigating the need for multiple testing correction. Simulations show that BBER can substantially improve sensitivity while maintaining high specificity. We also explore operating characteristics of the algorithm with respect to the structure of the data and discuss its strengths and weaknesses. Finally, we apply BBER to eQTL analysis using data from the Lung Genomic Research Consortium and find significant eQTL to be enriched for SNPs previously identified in genome-wide association studies of COPD.

Conclusions: BBER improves sensitivity and is suitable for discovery of genetic variants that may play roles in gene regulation. BBER can be applied to genome-wide eQTL analysis, and an implementation of BBER is available through Bioconductor Package BBER.

Introduction

Expression quantitative trait locus. (eQTL) analysis offers functional interpretation of disease-associated single nucleotide polymorphisms (SNPs) from genome-wide association studies (GWAS) [1-3]. In practice, eQTL analysis involves several million association tests between tens of thousands of quantitative gene expression measurements and millions of genetic markers, making multiple-testing correction both essential and challenging.

Genetic markers often possess a correlative structures due to linkage disequilibrium (LD), a process in which alleles at multiple, adjacent loci are inherited together as "LD blocks" and therefore correlate with one another. Gene expression profiles can also be correlated as genes in the same pathways may be co-regulated and therefore co-expressed. Direct application of the standard multiple-testing correction methods such as Benjamin-Hochberg false discovery rate (FDR) and Bonferroni procedure ignore this correlation structure and therefore are not appropriate for eQTL analysis as they can overestimate false discovery rates and over-correct p-values, both of which substantially reduce power to detect SNP-gene associations [4].

Several methods have been proposed to address the problem. Permutation testing is widely used in eQTL analysis [5, 6] because it preserves the correlation structure between genetic markers; heuristic or asymptotic approximation methods were developed to speed up the computationally intensive permutation test [7, 8]. However, because of its high computational cost, permutation testing is generally done only on SNPs; thus it does not correct for the correlation between gene expression measurements [4].

Chen et al. [9] and Kendziorski et al. [10] proposed mathematical frameworks to address both genotyping and gene expression correlation simultaneously. Chen et al. [9] developed a variance estimator for false discoveries and used weighted upper bound of false discovery proportion (wuFDP), with correlation between markers as weights, for false discovery control. Kendziorski et al. [10] proposed an empirical Bayes hierarchical mixture model that shares information across transcripts to determine a posterior probability that each transcript maps to each marker. While mathematically elegant, both of these approaches were developed for a small number of markers (a few hundred), making them less useful for a genome-wide eQTL analysis. Furthermore, Kendziorski's mathematical constructions also restrict the eQTL model to the hierarchical mixture model, disallowing adjustment of potential confounders.

Here we present BBER (bicluster-based error rate control), an algorithm that defines biclusters of correlated eQTL tests by hierarchical clustering and then applies a Benjamini-Hochberg FDR correction to the biclusters. BBER recognizes that association tests of SNPs in the same LD block yield highly correlated test statistics as they are effectively the same test; and the grouping reduces the effective number of independent tests

We developed and tested BBER using simulated data and then applied it to real data from 163 patients with and without chronic obstructive pulmonary disease (COPD) from the Lung Genomics Research Consortium (LGRC; http://lung-genomics.org/research).

Methods

Development of BBER procedure

The principle underlying BBER is that correlations due to the genome's haplotype structure and co-expression among genes in the same pathways effectively reduce the number of independent tests that should be performed. BBER takes advantage of this structure in the data and uses a modified Benjamini-Hochberg FDR adjustment (BH) based on the number of biclusters as opposed to the number of tests.

Simulation study

Following the approach described by Kendziorski et al. [4] and Chen et al. [11], we used a series of simulations to assess the performance of multiple testing correction procedures. The code for the simulation as well as the BBER procedure is available in the Bioconductor package BBER (URL). Genotypes of specified minor allele frequencies (MAF) were simulated as described by Emrich and Piedmonte [12] in which correlated binary data are simulated through dichotomizing multivariate normal vectors (Figure S5.1, top left panel). SNP-gene expression pairs were chosen at random to be the "ground truth" eQTLs (Figure S5.1, top middle panel) such that about half of the genes are regulated by one eQTL. The means of the gene expression profiles were calculated according to a genetic model (additive, recessive, or dominant), and the standard deviation is scaled linearly with the square root of the mean [13]. Given the mean and standard deviation, gene expression log intensities were drawn from the multivariate normal distribution (Figure S5.1, top right panel).

More formally, let a block correlation matrix $\mathbf{B}_{m,n}(\rho) = \mathbf{I}_m \otimes \mathbf{C}_n(\rho)$ where $\mathbf{I}_m$ is a $m \times m$ identity matrix and $\mathbf{C}_n(\rho)$ is a $n \times n$ correlation matrix with off-diagonal correlation ρ. Let G_{ip} denote the minor allele count of SNP p of Subject i, $G_{ip} \in [14]$, such that the matrix $\mathbf{G} = [G_{..}]$ has correlation $\text{Corr}(\mathbf{G}) = \mathrm{B}_{P,P}(r)$, and β denote the eQTL effect size. The indicator function $\mathbf{1}(p,q)$ is 1 if SNP-gene expression pair (SNP_p, Gene_q) is selected as an eQTL.

The mean and standard deviation of the gene expression q of Subject i are:

$$\mu_{iq} = \begin{cases} \beta \mathbf{1}(p,q) G_{ip} & \text{for additive model,} \\ \beta \mathbf{1}(p,q) \mathbf{1}(G_{ip} > 0) & \text{for dominant model,} \\ \beta \mathbf{1}(p,q) \mathbf{1}(G_{ip} = 2) & \text{for recessie model.} \end{cases}$$

and

$$\sigma_{iq} = 1 + \sqrt{\mu_{iq}}.$$

With the mean, standard deviation, and correlation matrix, we simulated gene expression profile for each subject by drawing from multivariate normal distribution. Thus, the log expression profile of Subject i is:

$$\log(Y_i) \sim \mathcal{N}(\mu_i, \mathbf{\Sigma}_i)$$

where $\mathbf{\Sigma}_i = \text{diag}(\sigma_i)\ \mathbf{P}(\rho)\ \text{diag}(\sigma_i)$ and $\mathbf{P}(\rho) = \mathbf{B}_{Q,Q}(\rho)$ is the expression correlation matrix.

The correlation matrices of genotype and gene expression data are of particular importance. They were specified to represent clusters of correlated markers with *intra*-cluster correlation r and ρ for genotype and gene expression data respectively; *inter*-cluster correlation is set to zero to represent independence between clusters. Evidence suggests the LD correlation r ranges between 0.5 and 1, while the co-expression correlation ρ is somewhat weaker. Therefore we used $r = 0.1$, 0.5, or 0.9 and $\rho = 0.5$ for our simulations.

Because each LD block and co-expression cluster is independent and identically distributed in our simulation, the numbers of LD blocks and gene clusters do not affect the results and were chosen so that the simulation can be completed in about three days on a 250-node cluster. Thus, we chose to simulate 10 LD blocks of 10 correlated SNPs and 5 gene expression clusters of 5 co-expressed genes, for number of subjects $N = 200$. All simulations were repeated 100 times. Also based on prior observations, the eQTL effect size β is around 2, so we simulated the eQTL data with β from 0.1 to 4 to capture the range of possible values.

Because each simulation was generated with a ground truth eQTL, we were able to calculate sensitivity and specificity and compare them across multiple testing correction methods. In the sensitivity analyses, we varied the key data generation parameters including LD correlation r, co-expression correlation ρ, and the eQTL regulation effect size β.

A linear model for eQTL analysis

We consider a model of eQTL analysis in the context of correlated data. For each individual ($i = 1, \ldots, N$), let $\mathbf{Y}_i = (Y_{i1}\ Y_{i2}\ \ldots\ Y_{iP})^{\mathrm{T}}$ be a vector of gene expression values from P different genes, and $\mathbf{G}_i = (G_{i1}\ G_{i2}\ \ldots\ G_{iQ})^{\mathrm{T}}$ be a genotype vector containing minor allele counts (0, 1, or 2) of Q SNPs. Let $\mathbf{X}_i$ be a covariate vector that contains 1 as the first element and may include environmental and demographic variables such as age, sex, and smoking status.

An association between gene expression $\mathbf{Y}_{\cdot p}$ and SNP $\mathbf{G}_{\cdot q}$ can be assessed through a generalized linear model:

$$E\left(Y_{ip} \middle| \mathbf{X}_i, \mathbf{G}_{iq}\right) = g(\mathbf{X}_i^T \alpha_{pq} + \mathbf{G}_{iq}^T \beta_{pq})$$

where α_{pq} is a vector of covariate effects, β_{pq} is a Q-dimensional vector of additive genetic effects, and g is a link function.

The corresponding PQ-dimensional vector of score statistics is:

$$\mathbf{U}_\beta = \sum_{i=0}^{N} \left(\mathbf{Y}_i - \widetilde{\mathbf{Y}}_i\right) \otimes \mathbf{G}_i$$

where $\widetilde{\mathbf{Y}}_i = g(\mathbf{X}_i^T \hat{\alpha}_{pq})$ is the vector of predicted gene expression trait values given covariates, under the assumption of no genetic association. And as shown by Conneely and Boehnke [7],

$$\mathbf{U}_\beta \sim N(\underline{0}, \mathbf{V}_\beta)$$

where $\mathbf{V}_\beta$ is estimable by $\mathbf{\Omega} \otimes (\mathbf{G}\mathbf{G}^T - \mathbf{G}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{G}^T)$, the Kronecker product of the sample covariance matrices of traits and genotypes $\mathbf{G} = (\mathbf{G}_1\ \mathbf{G}_2\ \ldots\ \mathbf{G}_N)$, conditioned on covariates $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2\ \ldots\ \mathbf{X}_N)$.

Following the definition above, the trait covariance matrix conditioned on X is

$$\mathbf{\Omega} = \sum_{i=0}^{N} (\mathbf{Y}_i - \widetilde{\mathbf{Y}}_i)(\mathbf{Y}_i - \widetilde{\mathbf{Y}}_i)^T.$$

Thus the normalized test statistics:

$$T_t = \frac{\mathbf{U}_{\beta,t}}{\sqrt{\mathbf{V}_{\beta,tt}}}.$$

And $\mathbf{T} \sim N(0, \mathbf{R})$ where the correlation matrix $\mathbf{R}$ can be estimated as a Kronecker product of the sample correlation matrices of gene expression traits and genotypes. Therefore, the correlation between two score statistics $\mathrm{Cor}(T_{p,q}, T_{p',q'})$ is a product of the correlation between the genotypes $\mathrm{Cor}(\mathbf{G}_q, \mathbf{G}_{q'})$ and the correlation between the gene expression profiles $\mathrm{Cor}(\mathbf{Y}_p, \mathbf{Y}_{p'})$.

Biclusters of SNPs and gene expression profiles

The analysis above suggests that the correlation between eQTL test statistics can be factorized into two independent components: correlation between SNPs and correlation between gene expression profiles. Thus, biclusters of eQTL tests can be defined as combinations of clusters of correlated SNPs (LD blocks) and clusters of correlated gene expression. More precisely, if

$$GE_k = \{p \mid \mathrm{Cor}(\mathbf{Y}_p, \mathbf{Y}_{p'}) > \rho, \text{ for all } p' \text{ in } GE_k\},$$

for some ρ, and for $k = 1, \ldots, K$,

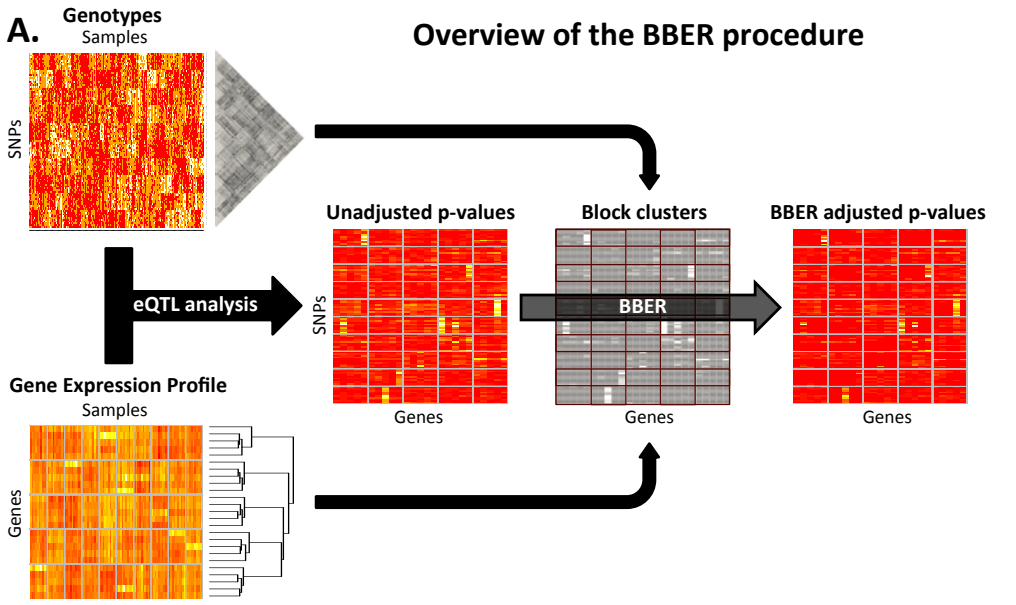
$$LD_l = \{q \mid \mathrm{Cor}(\mathbf{G}_q, \mathbf{G}_{q'}) > r, \text{ for all } q' \text{ in } LD_l\},$$

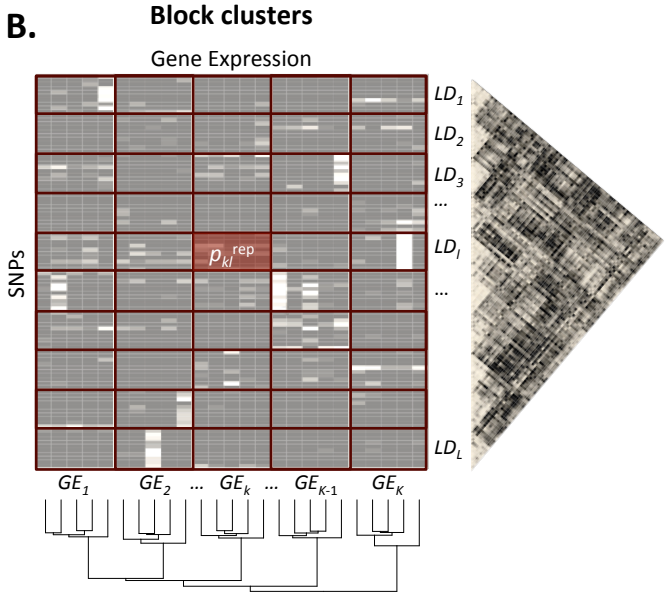
for some r, and for $l = 1, \ldots, L$, the corresponding bicluster is the Cartesian product:

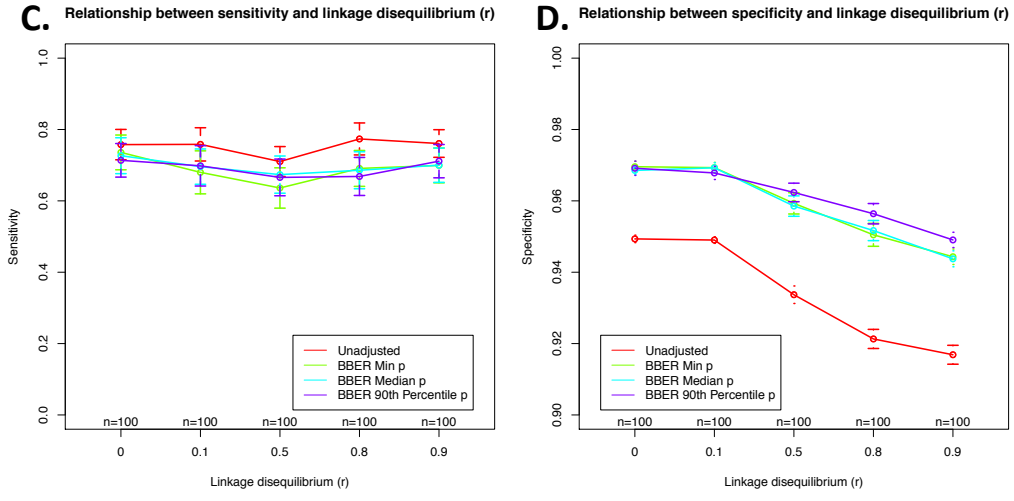
$$(GE_k, LD_l) = GE_k \times LD_l.$$

Practically, clustering of both SNPs and gene expression can be achieved through hierarchical clustering with absolute Pearson's correlation distance ($1 - |\mathrm{corr}|$) and complete linkage agglomeration method.

By cutting the clustering dendrogram at the height x, we can define clusters of with absolute intra-cluster correlation of $1 - x$. To maximize the inter-cluster distance (Figure S5.2), we chose to cut the dendrogram at $x = 0.4$. In the case of SNP clustering, there exist a number of haplotype block detection methods [5, 15], which take advantage of the biological fact that SNPs in LD tend to be in close proximity, typically within a few hundred thousand base pairs [16].


A.
Genotypes
Samples
SNPs
Overview of the BBER procedure
Unadjusted p-values
Block clusters
BBER adjusted p-values
eQTL analysis
SNPs
BBER
Genes
Genes
Genes
Gene Expression Profile
Samples
Genes


B.
Block clusters
Gene Expression
SNPs
LD1
LD2
LD3
...
LDl
...
LDL
GE1
GE2
...
GEk
...
GEK-1
GEK


C.
Relationship between sensitivity and linkage disequilibrium (r)
Sensitivity
Unadjusted
BBER Min p
BBER Median p
BBER 90th Percentile p
n=100
0
0.1
0.5
0.8
0.9
Linkage disequilibrium (r)
D.
Relationship between specificity and linkage disequilibrium (r)
Specificity
Linkage disequilibrium (r)

(continued)
Figure 5.1: Overview of BBER procedure. (A) eQTL analysis is performed by a standard method such as linear regression. LD blocks and gene expression clusters are identified independently from data, using hierarchical clustering with complete linkage and absolute Pearson's correlation distance. (B) A grid defined by the LD blocks and gene expression clusters constitutes biclusters. We select the 90th percentile p-value from each bicluster as the representative statistic to be ranked across all biclusters in the Benjamini-Hochberg procedure. We use the number and rank of the biclusters instead of the individual tests to obtain adjusted p-values. (C-D) 90th percentile of the p-values is the best bicluster representative statistics in terms of specificity, while all representative statistics achieve comparable sensitivity.

Multiple hypothesis testing adjustment

The Benjamini-Hochberg (BH) procedure [4] uses a false discovery rate (FDR) adjustment to correct for multiple hypothesis testing. The BH adjustment assumes that all p-values are independent. Suppose that V is the number of false discoveries, R is the number of rejected null hypotheses ("discoveries"), and S is the number of true positives. The FDR is given by:

$$\text{FDR} = \text{E}[V/R] = \text{E}[V/(V+S)]$$

Under the null hypothesis, the p-values are uniformly distributed, U(0,1), and V can be estimated by the rank of the p-value. Thus, FDR < α when $H_{(1)}$, $H_{(2)}$, …, $H_{(k)}$ are rejected for the largest integer k such that $q_{(k)} = p_{(k)}\, m/k < \alpha$, where m is the total number of tests.

BBER-adjusted P-value

After clustering eQTL tests into biclusters, we treat each bicluster as an independent test to be corrected for in the Benjamini-Hochberg FDR procedure. For each bicluster, defined by a pair of gene expression cluster GE_k and LD block LD_l (Figure 5.1A-B), we use the 90th percentile of the p-values from all individual eQTL tests $T \in \{T_{pq} | p \in GE_k, q \in LD_l\}$ to represent the

bicluster. That is, for a bicluster (GE_k, LD_l), the representative statistic is $p_{kl}^{\text{rep}} = P_{90}(p_{pq} | p \in GE_k, q \in LD_l)$. And p_{kl}^{rep} are then raked:

$$p_{(1)}^{\text{rep}} \leq p_{(2)}^{\text{rep}} \leq \cdots \leq p_{(KL)}^{\text{rep}}.$$

and if t is the rank of p_{kl}^{rep}, i.e. $p_{kl}^{\text{rep}} = p_{(t)}^{\text{rep}}$, then the BBER-adjusted FDR is:

$$p_{pq}^{\text{adj}} = \frac{KL}{t} p_{pq}$$

for all p in GE_k and q in LD_l.

Choice of the bicluster representative statistic

In BBER, a representative statistic is selected for each bicluster of correlated tests, and that choice is crucial to the performance of the method (Figure 5.1B). As true associations likely exhibit strong signal compared to other associations in the same bicluster, extreme statistics such as minimum p-values likely correspond to the true eQTL. However, minimum p-value could bias towards larger clusters, as they have a higher chance of containing extreme p-values. A robust statistic such as median could be more appropriate. Though robust to biases, median p-values might represent the true signal poorly while a higher percentile, such as the 90$^{\text{th}}$ percentile, might present a better compromise.

We used simulations to compare sensitivity and specificity of BBER under three different statistics: minimum p-value, median p-value, and 90$^{\text{th}}$ percentile of the p-values. As show in Figure 5.1C-D, simulations suggest 90$^{\text{th}}$ percentile p-value outperforms minimum and median p-values on specificity while maintaining comparable sensitivity. Based on these results, we decided to implement BBER using the 90$^{\text{th}}$ percentile p-value as the representative statistic.

Application to the Lung Genomics Research Consortium Data

We then applied BBER to data from a case-control study of chronic obstructive pulmonary disease (COPD). We retrieved clinical metadata and gene expression profiles from the Lung Genomics Research Consortium (LGRC) Data Portal (https://www.lung-genomics.org/research) and genotyping data from the same study (dbGaP accession number: phs000624.v1.p1). We selected samples having both gene expression and genotyping data. Gene expression data (collected using the Agilent Technologies G4112F array) contains information on 14,557 Ensembl genes. Following GCRMA normalization [17], we used surrogate variable analysis (SVA) [18] and found no batch effects.

Genotype data (collected using the Illumina Infinium HD Assay with Human Omni QuAd arrays) contains 2,443,179 markers. To control for genotyping errors, we removed any markers that failed Hardy-Weinberg equilibrium test ($p < 0.001$, 1,166 markers removed), were missing in more than 10% of samples (35,577 markers removed), or had minor allele frequency lower than 5% (1,131,328 markers removed). There were 1,288,792 SNPs remaining. Related subjects based on identity by descent were also removed. This quality control yields 126 COPD and 37 control samples for the eQTL analysis (86 males, 77 females, 163 total).

eQTL analysis

Using the LGRC data, we performed genome-wide expression quantitative trait loci (eQTL) analysis, which fits mRNA expression to the following model:

$$Gene\ Expression = \beta_0 + \beta_1 ADD + \beta_2 Pkyr + \beta_3 Age + \beta_4 Sex + \boldsymbol{\varepsilon}$$

where *ADD* is the allele dosage of a genotype (coded as 0, 1, and 2 for the number of minor alleles) and *Pkyr* is the number of pack years of smoking. We used PLINK v.1.07 to carry out the analysis with the command:

```
plink --bfile <input.genotype> --pheno <input.expression> --all-pheno --
    linear --pfilter 0.05 --adjust --covar <pkyrs.age.sex>
```

The p-value filter (pfilter flag) discarded all SNP-gene expression pairs with (unadjusted) p-values less than 0.05 as those would become insignificant after multiple testing adjustment and would not affect the outcome of the BBER procedure.

LD block identification

LD blocks were identified by the command "blocks" in PLINK v.1.07 [6]:

```
plink --bfile <input.genotype> --blocks --hap-freq
```

The algorithm follows the default procedure in Haploview [5].

Gene expression clustering

To obtain gene expression clusters, we used R to perform complete linkage hierarchical clustering with absolute Pearson's correlation distance, 1-|Corr|, on the expression profiles of the 14,557 genes represented on the Agilent array. We used absolute Pearson's correlation as gene pairs that are either negatively or positively correlated can often be shown to be co-regulated. We defined clusters by cutting the hierarchical clustering dendrogram at the height 0.4, which optimizes for the intra-cluster distance (Figure S5.2). This produced 8,288 clusters

with an average cluster size of 1.75 genes (max cluster size 81 genes). The small average cluster size reflects independence of the expression probes and is expected.

Enrichment of GWAS SNPs

We assessed enrichment of published COPD-related genome-wide associations in our eQTL SNPs using the database of publicly available GWAS (GWASdb; http://jjwanglab.org/gwasdb/, version 3, released 6/20/2013, accessed 10/28/2013). Filtering from a total of 170,385 SNPs in GWASdb using keywords "chronic obstructive pulmonary disease" and "emphysema," we have 178 unique SNPs from in 7 studies (see Supplemental Table S5.1); of these COPD-associated SNPs, 93 are present in the LGRC SNP set. Using LD structure in the LGRC samples, we inferred an additional 718 SNPs previously associated with COPD. We used Fisher's exact test to assess overrepresentation of the overlap between the GWAS SNPs and eQTL SNPs and used Wilcoxon rank sum test to show that BBER-adjusted p-values of GWAS SNPs are lower than those of non-GWAS SNPs. The enrichment shows that our eQTL results identify genetic variants relevant to COPD.

Results and Discussion

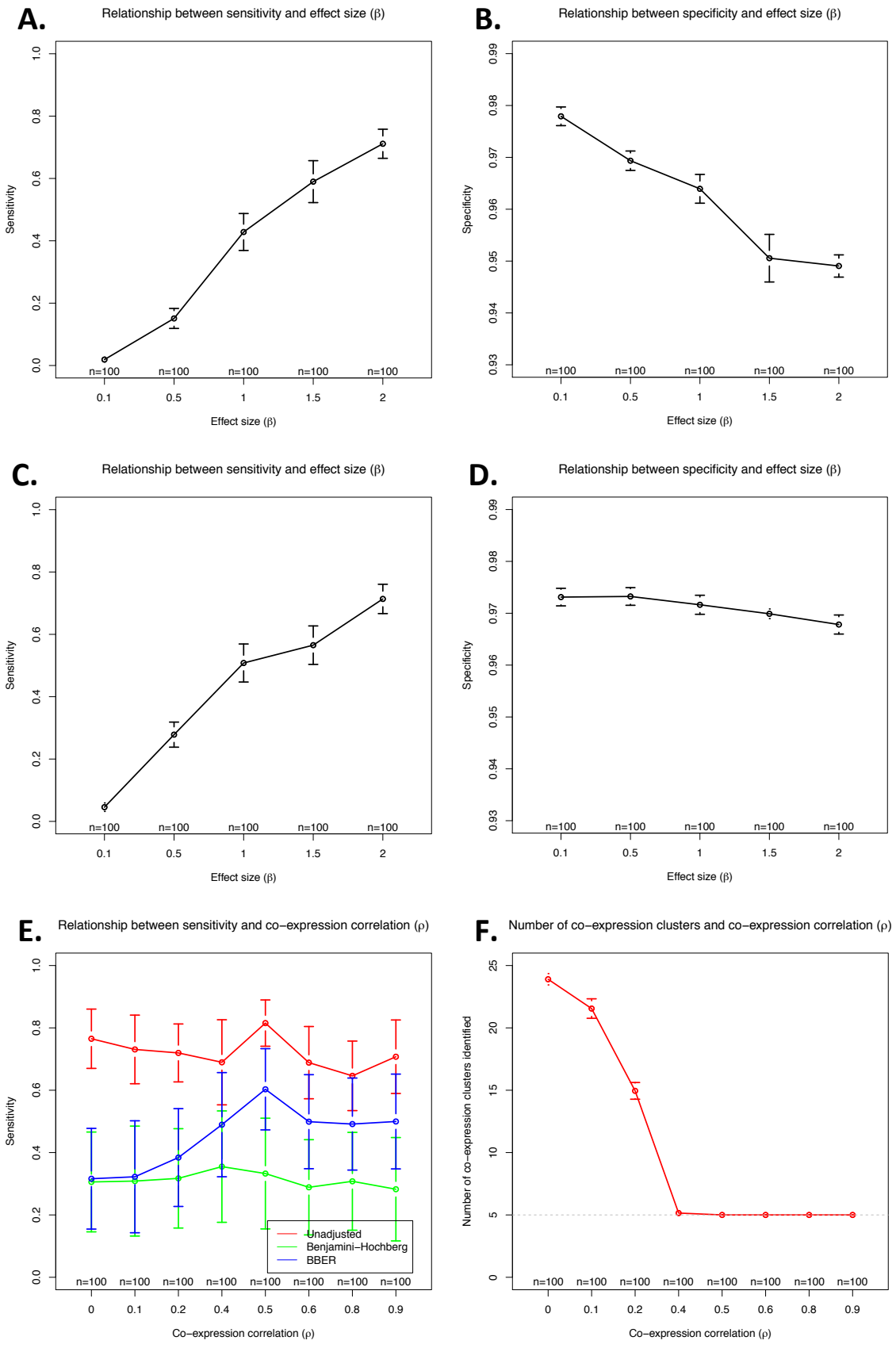
BBER improves specificity with little compromise on sensitivity

Compared to multiple-testing adjustment approaches like Benjamini-Hochberg (BH) and Bonferroni corrections, BBER limits over-correction by grouping dependent tests into blocks and only corrects for the number of the *de facto* independent blocks. Thus we expected BBER would improve on sensitivity while maintaining high specificity.

We demonstrated this using a series of simulations, following the approach described by Kendziorski et al. [4] and Chen et al. [11]. Although the simulation may not capture the complexities of a real dataset, it allows for an unbiased assessment the performance of BBER relative to other methods and helps explicate its operating characteristics.

The simulation parameters were chosen to represent a typical eQTL dataset (based on an empirical observation of the LGRC data; see Methods for dataset description): the average SNP MAF of 0.3, effect size $\beta = 2$, and the average intra-cluster expression correlation of 0.5. All three genetic models (additive, recessive, and dominant) and three levels of LD correlation ($r = 0.1$, 0.5, and 0.9) were considered. The eQTL analysis used a linear regression under the additive model assumption.

Results in Table 5.1 show reduced power of BH and Bonferroni approaches and highlight advantages of BBER in improving power. If we regard sensitivity achieved by the unadjusted cases as the maximal achievable sensitivity, using BH and Bonferroni procedures reduces sensitivity to detect additive effects by as much as 58.6% (from 0.76 to 0.31) and 66.4% (from 0.76 to 0.26), respectively (first row of Table 5.1). The reduction is more severe in the cases of the dominant and recessive models, possibly due to the lower baseline sensitivity. The sub-optimal sensitivity demonstrates how BH and Bonferroni procedures over-adjust the p-values due to the independence assumption. A major tradeoff of BBER is a slightly lower specificity compared to BH and Bonferroni. Depending on the application, this tradeoff can be worthwhile as BBER can improve sensitivity of the analysis while maintaining a reasonable specificity ($\geq$ 0.95).


A.
Relationship between sensitivity and effect size (β)
Sensitivity
Effect size (β)
n=100
B.
Relationship between specificity and effect size (β)
Specificity
Effect size (β)
n=100
C.
Relationship between sensitivity and effect size (β)
Sensitivity
Effect size (β)
n=100
D.
Relationship between specificity and effect size (β)
Specificity
Effect size (β)
n=100
E.
Relationship between sensitivity and co−expression correlation (ρ)
Sensitivity
Co−expression correlation (ρ)
Unadjusted
Benjamini−Hochberg
BBER
n=100
F.
Number of co−expression clusters and co−expression correlation (ρ)
Number of co−expression clusters identified
Co−expression correlation (ρ)
n=100

(continued)

Figure 5.2: Sensitivity analyses. All simulations were produced with parameters: $N = 200$, MAF of 0.3, effect size $\beta = 2$, LD correlation $r = 0.9$, and co-expression correlation $\rho = 0.5$, unless otherwise noted. Each simulation is repeated 100 times ($n = 100$) (A-B) Sensitivity and specificity of BBER at various eQTL effect size β. Sensitivity increases as the effect size increases. For correlated markers ($r = 0.9$, $\rho = 0.5$), the increase in sensitivity is accompanied by decrease in specificity. (C-D) This decrease in specificity is a result of strong correlation between genetic markers. When markers are weakly correlated ($r = 0.1$), the specificity does not decrease as a function of the effect size β. (E) Sensitivity of BBER is similar to BH at low ρ correlations and to the unadjusted case at high ρ correlations. (F) The bimodality behavior of BBER can be explained by convergence of the number of clusters to the "true" number ($k = 5$), which coincides with the convergence of BBER sensitivity toward the unadjusted case.

		BBER		**Unadjusted**		**BH**		**Bonferroni**	
Model	*r*	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
Additive	0.1	0.698 (0.285)	0.968 (0.009)	0.758 (0.235)	0.949 (0.005)	0.314 (0.384)	1.000 (0.000)	0.255 (0.335)	1.000 (0.000)
	0.5	0.666 (0.259)	0.962 (0.013)	0.710 (0.211)	0.934 (0.012)	0.333 (0.355)	0.994 (0.011)	0.286 (0.318)	0.998 (0.005)
	0.9	0.711 (0.235)	0.949 (0.011)	0.761 (0.195)	0.917 (0.013)	0.416 (0.360)	0.983 (0.018)	0.321 (0.325)	0.989 (0.012)
Dominant	0.1	0.537 (0.293)	0.970 (0.007)	0.600 (0.283)	0.948 (0.006)	0.211 (0.280)	1.000 (0.000)	0.167 (0.215)	1.000 (0.000)
	0.5	0.506 (0.317)	0.968 (0.009)	0.607 (0.273)	0.939 (0.010)	0.212 (0.297)	0.998 (0.004)	0.175 (0.234)	1.000 (0.001)
	0.9	0.590 (0.274)	0.959 (0.014)	0.633 (0.236)	0.929 (0.019)	0.289 (0.324)	0.990 (0.014)	0.167 (0.227)	0.995 (0.009)
Recessive	0.1	0.388 (0.202)	0.968 (0.009)	0.514 (0.196)	0.949 (0.006)	0.061 (0.103)	1.000 (0.000)	0.054 (0.096)	1.000 (0.000)
	0.5	0.381 (0.257)	0.974 (0.006)	0.552 (0.288)	0.940 (0.012)	0.094 (0.152)	0.999 (0.001)	0.065 (0.100)	1.000 (0.000)
	0.9	0.387 (0.267)	0.965 (0.011)	0.545 (0.218)	0.932 (0.016)	0.093 (0.202)	0.996 (0.008)	0.056 (0.113)	0.999 (0.002)

Table 5.1: Sensitivity and specificity of BBER and other multiple testing adjustments. The values shown are mean with standard deviation in parenthesis.

Sensitivity Analysis

To understand the operating characteristics of BBER, we observed how specificity and sensitivity change in response to varying key data simulation parameters. Of the parameters, two are particularly important: eQTL effect size and correlation between genetic/gene expression markers. The eQTL effect size is important as it directly affects the eQTL linear model and influences the specificity-sensitivity tradeoff. As BBER relies on the ability to group together correlated markers, the correlation between markers has a profound impact of the quality and accuracy of clustering, thus influencing BBER.

eQTL Effect Size

In a typical statistical test, the power (sensitivity) of the test increases as the effect size increases while specificity, which is a function of the false positive rate, remains constant and independent of the sensitivity. However, when tests are correlated, as is the case for genetic data, a greater effect size could bias the p-value distribution, leading to an elevated false positive rate. We demonstrated this by varying the eQTL effect size β from 0.1 to 2 (values chosen based on empirical observations of the LGRC data). When the markers are correlated ($r = 0.9$, $\rho = 0.5$), power increases as a function of the effect size (Figure 5.2A) as expected, but we observed a slight decrease in specificity as a tradeoff (Figure 5.2B). This reduction in specificity is a result of the high correlation between genetic markers ($r = 0.9$) as this decrease in specificity is much more subtle when the genetic markers are only weakly correlated ($r = 0.1$, Figure 5.2D).

Clustering and BBER

The strength of the correlation between genetic and gene expression markers influences the performance of BBER as it determines the number of biclusters to correct for. At one

extreme, if markers are highly correlated, and the clustering algorithm groups all tests into one single cluster, there will be no adjustment, and BBER is equivalent to the unadjusted case. On the other hand, if markers are all independent, and each test is its own cluster, BBER is equivalent to BH. In an ideal case, BBER should strike the balance by appropriately grouping together correlated tests. Thus intuitively weakly correlated markers should result in BH-like results, and conversely, highly correlated markers should result in unadjusted case-like behaviors.

To test this, we simulated gene expression to have different levels of correlation (0-0.9) while keeping the correlation between genotypes low, at $r = 0.1$. As expected, at low expression correlation ($\rho < 0.4$), sensitivity of BBER resembles that of BH and at high correlation ($\rho \geq 0.4$), it resembles the unadjusted case (Figure 5.2E). We further validated the intuition that hierarchical clustering, which forms a basis for BBER algorithm, is adaptive to the strength of intra-cluster correlation and identifies the correct number of clusters. As show in Figure 5.2F, when gene expression correlations are low ($\rho < 0.4$), the number of clusters identified is high, reflecting the weak clustering. But as expression correlation ρ increases, the identified number of clusters converges to the "true" number of clusters ($k = 5$), and the sensitivity of BBER starts to resemble that of the unadjusted case.

It is not coincidental that 0.4 is the point at which BBER "switches" from BH-like sensitivity to unadjusted case-like sensitivity. In the BBER procedure, 0.4 was used to define clusters in hierarchical clustering and we selected 0.4 to maximize inter-cluster distance (Figure S5.2). This cutoff can and should be changed based on the specific characteristics of a dataset.

Validating BBER specificity with biological data

To corroborate the results of the simulation study, we used real data to construct a negative control dataset by permuting the links between genotype and gene expression data from Lung Genomics Research Consortium (LGRC). We accomplished this by label-swapping gene expression data, as described previously [19]. Because the permutation procedure breaks genotype-gene expression association, any positive calls can be regarded as false positives. We used this permuted dataset to assess specificity of BBER as well as other multiple testing adjustment methods. As shown in Table 5.2, specificity from various methods agreed with our previous simulation results and confirmed that BBER was able to improve specificity of the eQTL analysis.

Method	Specificity mean (sd)
BBER	0.9753 (0.0086)
Unadjusted	0.9508 (0.0047)
Permutation	0.9522 (0.0052)
Bonferroni	1.0000 (0.0008)
BH	1.0000 (0.0011)

Table 5.2: Specificity of BBER and other methods, estimated from permuted LGRC data.

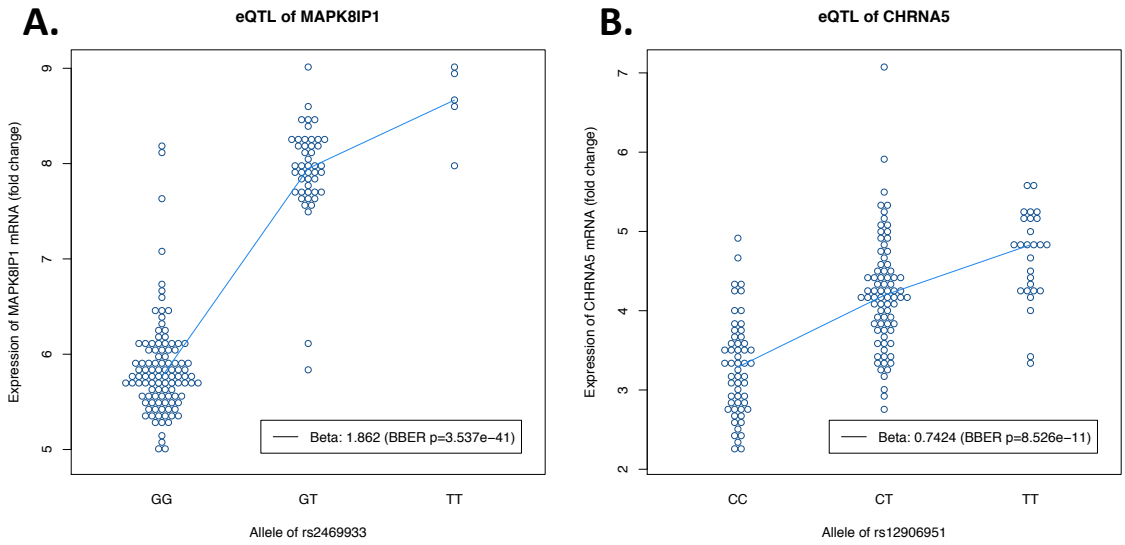


Figure 5.3: Examples of eQTL identified by BBER. Each point represents genotype and gene expression of one sample, and the line connects median expression of the genotype groups. (A) The eQTL associating the SNP rs16940665 with the expression of MAPK8IP1 gene is among the strongest eQTL. (B) The SNP rs12906951 is a *cis*-eQTL SNP of CHRNA5, an important COPD-associated gene [20].

BBER identified COPD-associated variants found in a large GWAS

We used BBER to analyze data from Lung Genomics Research Consortium (LGRC; lung-genomics.org). There are 163 LGRC subjects with both gene expression and genotype data: 126 COPD patients and 37 controls; the subjects are all Caucasians. After normalization and data processing (see Methods), we performed eQTL analysis using a linear model adjusted for three standard potential confounders for lung function modeling: pack years of smoking, age, and sex, and corrected for multiple-hypothesis testing using BBER. LD blocks were defined by PLINK [6], and gene expression was clustered using complete-linkage hierarchical clustering with absolute Pearson's correlation distance. We used a cutoff of 0.4 to define clusters in the dendrogram, which optimizes for the inter-cluster distance (see Figure S5.2). The size distributions of the LD blocks and gene clusters are given in Figure S5.3. Overall, there were

373,708 LD blocks (average size 3.4 SNPs/block) and 8,288 gene clusters (average size 1.75 genes/cluster).

There are 62,739 significant eQTL at the level 0.001, a conservative cutoff which gives high-confident eQTL calls. Of those, 5,991 eQTL are *cis*-acting, (SNPs within 1Mb from genes' transcription start sites). Consistent with conventional wisdom [3], the *cis*-eQTL are more statistically significant, as the BBER-adjusted FDRs are lower compared to the *trans*-eQTL (Wilcoxon p-value < 2.2×10^{-16}). In total, there are 36,364 SNPs and 9,614 genes with significant eQTL. These comprise about 2.82% of all SNPs and a substantial portion (63.37%) of the genes in the dataset. eQTL with strongest BBER-adjusted FDRs include *trans*-eQTL of MAPK8IP1 on Chromosome 17q21.31 (Figure 5.3A, BBER FDR=3.5×10^{-41}), *cis*-eQTL of OR7D2 (BBER FDR=1.1×10^{-80}), LDHC (BBER FDR=3.0×10^{-44}), PSPH (BBER FDR=3.0×10^{-41}), LRRC14 (BBER FDR=5.2×10^{-41}), AMFR (BBER FDR=2.4×10^{-37}), and SLC5A11 (BBER FDR=1.1×10^{-36}). The list of top eQTL is given in Table S5.1.

Because disease-associated variants have been found to often be regulatory [21], the COPD-associated SNPs are likely to be in eQTL, and so one would expect enrichment of variants associated with COPD. From GWASdb, a database of published genome wide association studies [22], we identified 178 COPD-associated SNPS from 7 COPD association studies. Of those, 93 SNPs are present in the LGRC genotyping arrays (Supplemental Table S5.1).

Since SNPs in GWASdb are mostly tag SNPs, a representative subset for other SNPs in the haplotype blocks, the overlap between GWASdb and the LGRC SNP set under-estimates the

number of COPD-associated SNPs in the LGRC dataset. Hence, we inferred association of additional 718 SNPs that are in LD with the 93 COPD-associated SNPs. Of these SNPs, 42 are in our significant eQTL list, representing a 1.6-fold enrichment over the genome-wide background (Fisher's exact p-value = 0.002) (Supplemental Table S5.3). Moreover, these COPD-associated eQTL SNPs tend to have strong BBER FDR compared to the non-COPD-associated eQTL (Wilcoxon p-value < 1.27×10^{-8}).

In addition to the overall enrichment of COPD GWAS hits, a number of well-replicated COPD SNPs are present. For example, SNPs in an important COPD-associated region on Chromosome 15q25.1 spanning CHRNA3, CHRNA5, and IREB2 [20] are present in our results as *cis*-eQTL of CHRNA5 gene expression (Figure 5.3B). A set of COPD-associated SNPs on Chromosome 6q21.3 includes *cis*-eQTL of PSORS1C1 and POU5F1. The eQTL SNP of POU5F1 (rs2074488) has been associated with circulating level of surfactant protein D (SP-D), which is linked exacerbations of COPD [23].

SNPs in this 6q21.3 region are also associated with COPD susceptibility in a previous eQTL analysis of sputum gene expression [24]. Another important COPD gene is hedgehog interacting protein (HHIP), for which multiple SNPs have been strongly associated with lung function and COPD [20, 25-29]. Ten significant *trans*-eQTL of HHIP gene are found in our analysis (Supplemental Table S5.4). The enrichment of COPD-associated SNPs in our eQTL results and the ability of BEBR to identify eQTL for important COPD genes CHRNA5 and HHIP suggest that BBER can detect regulatory signals relevant to COPD.

Conclusions

Standard multiple testing adjustment approaches for eQTL analysis make assumptions about independence of genotypes and gene expression probes that do not reflect the structure of relationships in real data, potentially resulting in overestimation of the number of independent tests and therefore greatly reducing the power. We developed BBER (bicluster-based error rate control) as an extension of the standard Benjamini-Hochberg (BH) by reducing the effective number of tests from the number of individual tests to the number of biclusters of correlated tests. Based on a series of simulation studies, BBER is able to improve sensitivity over Benjamini-Hochberg and Bonferroni procedures while maintaining the improved specificity. The method is scalable and can be applied to a genome-scale eQTL analysis. BBER has been implemented in R and is available as a Bioconductor package, BBER.

Sensitivity analysis demonstrated that BBER can be characterized as an adaptive compromise between two approaches: BH and no adjustment. BBER is more similar to BH when markers are weakly correlated and is more similar to no adjustment when markers are highly correlated. In reality, the strength of marker correlation is highly variable. The main advantage of BBER is that it is a data-driven way to strike the balance between these two commonly used approaches.

Based on simulation studies, BBER was able to improve sensitivity substantially over BH and Bonferroni procedures but with some tradeoff on the specificity. Thus, BBER is most appropriate for hypothesis generation rather than a confirmatory study. In an exploratory

study, the eQTL results can corroborate other lines of evidence such as GWAS and differential expression studies and help prioritize disease-associated variants.

In our application of BBER to LGRC data, we found our eQTL SNPs to be enriched for COPD-associated SNPs from several GWAS. Intersection of the two lines of evidence also sheds light on important COPD-associated regions such as Chromosome 15q25.1 (CHRNA5) [20] and Chromosome 4q31 (HHIP) [20, 25-29] and highlights a promising region on Chromosome 6q21.3, which has recently been associated with COPD [24].

Although BBER derives motivation from eQTL analysis, the method is generalizable to any type of multidimensional association studies. For example, correlation between DNA methylation and gene expression, sometimes called mQTL analysis, suffers the same problem of high dimensionality and highly correlated markers. As genomic technologies continue to advance, we will see increasing diversity and volume of high-dimensional data. There will be a need to integrate multiple data types at higher dimensionality and with possibly stronger correlation among markers. A method such as BEBR that accounts for such correlation may prove increasingly useful.

Chapter 5 Bibliography

1. Bochner BR: New technologies to assess genotype-phenotype relationships. *Nature reviews Genetics* 2003, 4(4):309-314.

2. Rockman MV, Kruglyak L: Genetics of global gene expression. *Nature reviews Genetics* 2006, 7(11):862-872.

3. Michaelson JJ, Loguercio S, Beyer A: Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* 2009, 48(3):265-276.

4. Kendziorski C, Wang P: A review of statistical methods for expression quantitative trait loci mapping. *Mammalian genome : official journal of the International Mammalian Genome Society* 2006, 17(6):509-517.

5. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005, 21(2):263-265.

6. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 2007, 81(3):559-575.

7. Conneely KN, Boehnke M: So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *American journal of human genetics* 2007, 81(6):1158-1168.

8. Zhang X, Huang S, Sun W, Wang W: Rapid and robust resampling-based multiple-testing correction with application in a genome-wide expression quantitative trait loci study. *Genetics* 2012, 190(4):1511-1520.

9. Chen L, Tong T, Zhao H: Considering dependence among genes and markers for false discovery control in eQTL mapping. *Bioinformatics* 2008, 24(18):2015-2022.

10. Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD: Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* 2006, 62(1):19-27.

11. Chen M, Kendziorski C: A statistical framework for expression quantitative trait loci mapping. *Genetics* 2007, 177(2):761-771.

12. Emrich LJ, Piedmonte MR: A Method for Generating High-Dimensional Multivariate Binary Variates. *Am Stat* 1991, 45(4):302-304.

13. Hornung G, Bar-Ziv R, Rosin D, Tokuriki N, Tawfik DS, Oren M, Barkai N: Noise-mean relationship in mutated promoters. *Genome research* 2012, 22(12):2409-2417.

14. Parmigiani G, Boca S, Ding J, Trippa L: Statistical tools and R software for cancer driver probabilities. *Methods Mol Biol* 2014, 1101:113-134.

15. A haplotype map of the human genome. *Nature* 2005, 437(7063):1299-1320.

16. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nature genetics* 2001, 29(2):229-232.

17. F WZIRGRM-MFS: A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* 2004, 99:909.

18. Leek JT, Storey JD: Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* 2007, 3(9):1724-1735.

19. Churchill GA, Doerge RW: Empirical threshold values for quantitative trait mapping. *Genetics* 1994, 138(3):963-971.

20. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, Feng S, Hersh CP, Bakke P, Gulsvik A *et al*: A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS genetics* 2009, 5(3):e1000421.

21. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Kelsoe JR, O'Donovan MC, Furberg H, Schork NJ *et al*: All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS genetics* 2013, 9(4):e1003449.

22. Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, Wong MP, Sham PC, Chanock SJ, Wang J: GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic acids research* 2012, 40(Database issue):D1047-1054.

23. Lomas DA, Silverman EK, Edwards LD, Locantore NW, Miller BE, Horstman DH, Tal-Singer R: Serum surfactant protein D is steroid sensitive and associated with exacerbations of COPD. *The European respiratory journal* 2009, 34(1):95-102.

24. Qiu W, Cho MH, Riley JH, Anderson WH, Singh D, Bakke P, Gulsvik A, Litonjua AA, Lomas DA, Crapo JD *et al*: Genetics of sputum gene expression in chronic obstructive pulmonary disease. *PloS one* 2011, 6(9):e24395.

25. Zhou X, Baron RM, Hardin M, Cho MH, Zielinski J, Hawrylkiewicz I, Sliwinski P, Hersh CP, Mancini JD, Lu K *et al*: Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Human molecular genetics* 2012, 21(6):1325-1335.

26. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, Franceschini N, van Durme YM, Chen TH, Barr RG *et al*: Meta-analyses of genome-wide association

studies identify multiple loci associated with pulmonary function. *Nature genetics* 2010, 42(1):45-52.

27. Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, Obeidat M, Zhao JH, Ramasamy A, Zhai G, Vitart V *et al*: Genome-wide association study identifies five loci associated with lung function. *Nature genetics* 2010, 42(1):36-44.

28. Soler Artigas M, Wain LV, Repapi E, Obeidat M, Sayers I, Burton PR, Johnson T, Zhao JH, Albrecht E, Dominiczak AF *et al*: Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function. *American journal of respiratory and critical care medicine* 2011, 184(7):786-795.

29. Van Durme YM, Eijgelsheim M, Joos GF, Hofman A, Uitterlinden AG, Brusselle GG, Stricker BH: Hedgehog-interacting protein is a COPD susceptibility gene: the Rotterdam Study. *The European respiratory journal* 2010, 36(1):89-95.

Chapter 6: Conclusion

The clinic of the future

In 2011, two leading scientists and technologists Stephen Friend and Trey Ideker wrote a commentary in *Nature Biotechnology* [1] envisioning "the clinic of the future." In their vision, future routine doctor visit will involve collecting not only the standard seven biometrics (sex, age, height, weight, temperature, pulse rate, and blood pressure) but also bio-specimens used to monitor patient's molecular phenotypes such as buccal swab for full genome sequence, stool samples for microbiome composition in the gastrointestinal track, and blood or urine samples for mRNA, miRNA, proteome, and metabolome profiles. All of the information will be integrated with the patient's electronic health record, and the new information augments the history gathered in previous visits, constructing a holistic portrayal of the patient's health over his/her lifetime.

Like many technological visions of the past, "the clinic of the future" may not manifest itself as imagined by Friend and Ideker. Think back to the vision for Artificial Intelligence in the 1980s. Early success of chess-playing computers such as IBM Deep Blue inspired imagination of humanoid robots living among human, like those portrayed in the movie *Terminator*. Instead, the last few decades have seen artificial intelligence manifesting itself as "smart" augmentation to the existing technologies, for example book, movie, and song recommendation systems (e.g. Amazon, Netflix, and Pandora) and context-aware search engine (e.g. Google). How genomic technology will eventually integrate itself into the medical practice

might be analogous. Instead of requiring an all-new way of seeing and treating patients, genomics will build upon the current time-tested clinical practice. We may not be collecting blood or stool samples for all patients at every visit, but will for patients with related conditions, such as inflammatory bowel disease and diabetes. Family history is already routinely used as a line of diagnostic evidence for certain conditions. Genome sequencing would simply be the logical next step.

A road toward personalized medicine

Regardless of how "personalized medicine" will manifest, the future points toward an increasing use of genomic technologies. Major research hospitals are now starting to collect bio-specimens for routine genomic screens. For example, in 2011 Brigham and Women's Hospital and Dana-Farber Cancer Institute launched Profile project[1], which aims to genotype a large number of patients and make the data available to researchers in the system. Similar efforts are also being made across the county and the world, for instance VA Healthcare's Million Veteran Program and UK10K Program in the United Kingdom. With the potential influx of information, regardless of how the clinic of the future would shape up to be, three areas of advancement are required to advance our genomic future: technology development, links between disease and markers, and integrative analysis approaches.

[1] http://www.dana-farber.org/Research/Featured-Research/Profile-Somatic-Genotyping-Study.aspx

Technology Development

In order for a genomic technology to be useful in the clinical setting, it requires "clinical-grade" sensitivity and specificity and a reasonable turn-around time. A number of technologies routinely used in research do not match one or more of those requirements. The limiting factors are partly the biochemical assays themselves but perhaps to a greater extent the lack of bioinformatics tools to process and make sense of the data. While the sequencing technology has advanced at a rapid pace over the last decade, its full power can only be realized when coupled with well-reasoned and appropriately contextualized computational tools.

Exome sequencing is one of the applications of next generation sequencing that has been used in the clinical setting. Following the sensational success at the Medical College of Wisconsin [2], a number of commercial and academic labs — including Ambry Genetics; University of California, Los Angeles; Washington University's School of Medicine; and Cincinnati Children's Hospital — have recently begun offering clinical sequencing services. Because of its cost-effectiveness, exome sequencing has become the technology of choice.

ExomeCNV, described in Chapter 2, is an example of a tool used in clinical exome sequencing. Since its publication in 2011, ExomeCNV has been used in several studies and cited more than 60 times (Google Scholar, accessed September, 2013). More importantly, it has been used to supplement clinical exome sequencing (personal communication, Dr. Hane Lee) at UCLA Clinical Genomics Center. Although the tool was developed for research purposes, oversight by Genomics Data Board, which includes Board Certified Pathologists, Molecular

Geneticists, Molecular Cytogeneticists, Clinical Geneticists, Genetic Counselors, and Informatics Specialists, helps improve its utility and reliability.

Associating markers and disease

Genomics technologies help aid discovery of variants and measure molecular activities in a system, but they are only useful when put in a context of traits and disease. Most of the variants in the human genome are benign polymorphisms, and most of the expressed genes are housekeeping genes. Genome-wide association study and differential expression analysis are ways to connect the genomic information to clinical phenotypes.

Because of the high dimensionality and possibly small effect size, statistical techniques that optimize for the technology and biological signal are crucial. Several major findings in genomics are made possible by the unique combination of data, statistics, and clinical understanding. For example, in 2000, Perou et al. presented a seminal paper in breast cancer [3] that combines a groundbreaking microarray technology, a hierarchical clustering technique, and a deep understanding of cancer biology. The paper created a deep impact not only by clarifying four major subtypes of breast cancers but also by popularizing the use of microarray and hierarchical clustering.

In a similar manner, our work in Chapter 4 has combined data, statistics, and clinical insights to elucidate sexual dimorphism in COPD. Leveraging mRNA expression from Lung Genomics Research Consortium, Chapter 4 describes a technique to link differential gene expression to COPD in a sex-specific manner. A careful consideration of samples' disease classification based on a deep clinical insight has proven crucial, as without the clinical

adjudication the sample heterogeneity in our data would obscure what's already very subtle signals. Thus the success of a marker-disease association study depends on a close collaboration between technologists, bioinformaticians, and clinicians.

Integrative approach to personalized medicine

The power of genomics is derived from the versatility of genomics technologies. With a few common tools, we can interrogate multiple aspects of a biological system. For example next-generation sequencing can be applied to study gene function and regulation, discovering new variants, and sequencing genomes of new organisms *de novo* (Figure 1). Because complex diseases likely involve more than one type of variants, several studies collect multiple types of genomic information on the same set of patients. The Cancer Genome Atlas (cancergenome.nih.gov), one of the largest collaborative research project supported by National Cancer Institute, assays and makes available several types of genomic data on thousands of cancer samples, including mRNA expression, miRNA expression, protein expression, single nucleotide variations, copy number variations, RNA-seq, and DNA methylation. Lung Genomics Research Consortium (www.lung-genomics.org) is the largest and most comprehensive genomic survey for Chronic Obstructive Pulmonary Disease (COPD), profiles genotypes, mRNA expression, miRNA expression, RNA-seq, and DNA methylation of several hundreds of patients.

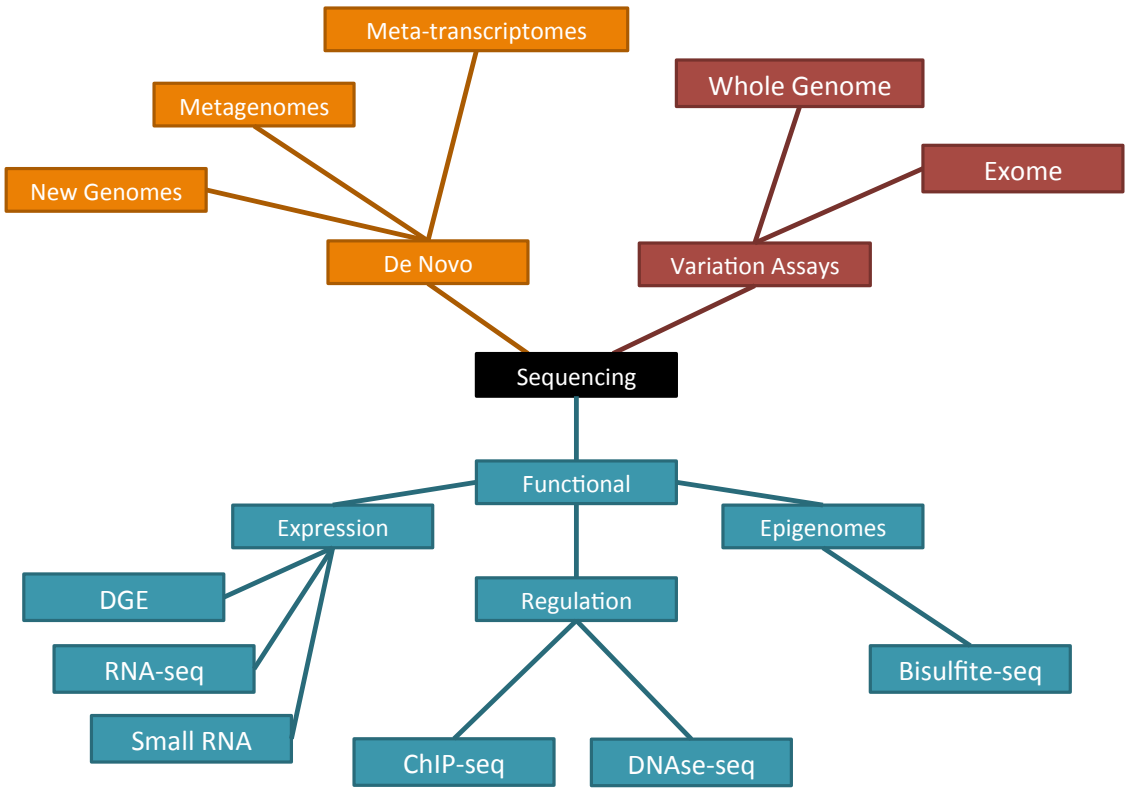
Each type of genomic data comes with a unique set of challenge. For example, in Chapter 3, we discuss challenges of studying co-occurrence and co-exclusion relationships among commensal microbes in the healthy human microbiome. There, we employ an ensemble

strategy to study various correlational patterns and develop a novel method called ReBoot to mitigate spurious correlation arising from compositionality of microbial abundance data.

Inasmuch as processing one type of data is challenging, much more so is integrating across data types. In Chapter 4 we corroborate our sexually dimorphic and COPD differential genes with DNA methylation, ChIP-seq binding sites of sex hormones, and genetic regulation of expression (eQTL). The eQTL analysis is our first step to integrate genotyping and gene expression data. We found that after adjusting for multiple hypothesis testing, the eQTL analysis was grossly underpowered. This served as a motivation for Chapter 5, which introduces BBER, a bicluster-based multiple hypothesis correction procedure. BBER extends the current standard Benjamini-Hochberg procedure by taking into account the profound correlation structure between co-expressed genes and within LD blocks. By taking advantage of the basic characteristics of gene expression and genotypes, we were able to improve sensitivity of eQTL analysis considerably.

Though challenging, integrating multiple types of data will be an important step forward in genomics. Already many studies have reported success in integrating information across platforms. The Cancer Genome Atlas Network has recently published comprehensive molecular portraits of human breast tumors, which combine copy number variation, DNA methylation, exome sequences, mRNA expression, miRNA expression, and protein expression to describe detailed molecular features of breast cancer subtypes [4]. More sophisticated integrative approaches such as the network model start to surface and show great promise [5]. There is a newfound appreciation for a more holistic view of a biological system, with an

emergence of Systems Biology [6] and Network Medicine [7], and integrative genomics is going to play an increasingly important role in pushing this frontier.



Reproduce from: http://finchtalk.geospiza.com/2010/09/genomics-genealogy.html

Figure 6.1: Applications of sequencing technology. Next-generation sequencing serves as a common platform to interrogate genetic map and molecular activities of a biological system.

Epilogue

The advancement in genomics insinuates not only the data integration but also the integration of disciplines. Because of its far-reaching field of applications, genomics touches and draws expertise from virtually all areas of biology, computer science, and statistics. A close collaboration between bioinformaticians and bench scientists creates a virtuous cycle of

hypothesis generation and experimental validation. Unbiased and high-throughput nature of genomic surveys can accelerate discovery of new biomarkers that can be developed into diagnostics and treatments. But without collaborators with the capability to validate the biomarkers and translate them to clinical uses, computational predictions would remain mere hypotheses.

To create a productive trans-disciplinary collaboration, scientists from both sides have to learn to appreciate and speak each other's language. For computational scientists, the ability to put a statistical method in a biological and experimental context can help improve the accuracy and more importantly the sensibility of the results. For experimental scientists, the ability to appreciate and intuitively evaluate different statistical models is key to choosing the most promising hypotheses to pursue.

With the whole world moving toward more open sharing of information, a phenomenon initiated by social media and accelerated by Meaningful Use and improvement on electronic health records, clinical information is going to become more connected and available. Making sense of all the information available on a patient's record is going to require a well-designed informatics engine that represents and summarizes the information effectively. More importantly, the clinic of the future will require a collaborative network of experts who help put context to data—to translate from data to information, and from information to understanding.

Chapter 6 Bibliography

1. Friend SH, Ideker T (2011) POINT: Are We Prepared For the Future Doctor Visit? Nature Biotechnology 29: 215-218.

2. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, et al. (2011) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 13: 255-262.

3. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. (2000) Molecular portraits of human breast tumours. Nature 406: 747-752.

4. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, et al. (2012) Comprehensive molecular portraits of human breast tumours. Nature 490: 61-70.

5. Glass K, Huttenhower C, Quackenbush J, Yuan GC (2013) Passing Messages between Biological Networks to Refine Predicted Interactions. Plos One 8.

6. Chuang HY, Hofree M, Ideker T (2010) A Decade of Systems Biology. Annual Review of Cell and Developmental Biology, Vol 26 26: 721-744.

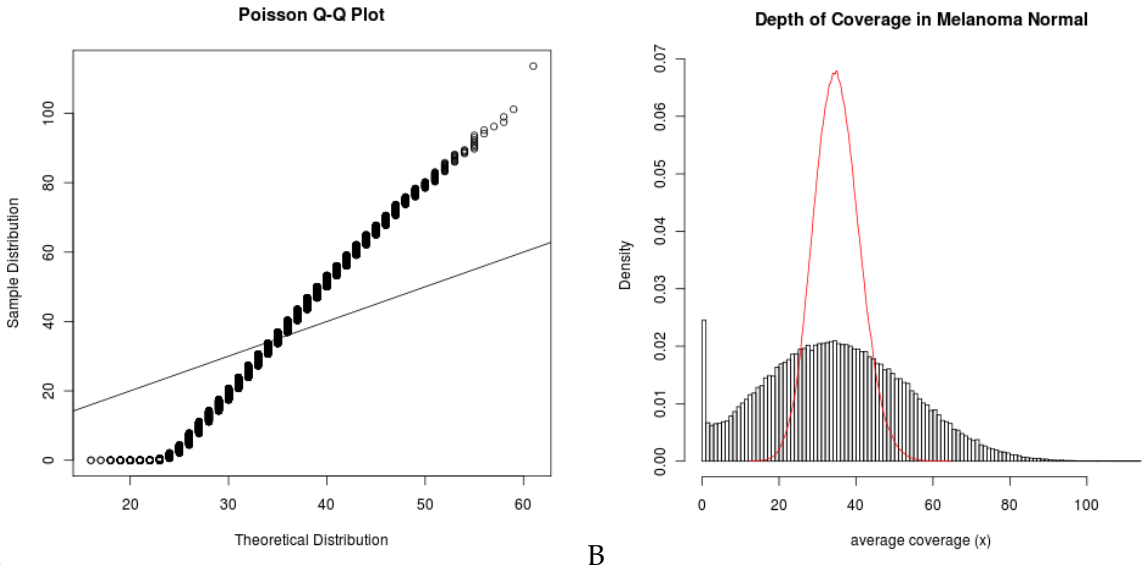
7. Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nature Reviews Genetics 12: 56-68.

Appendix 2A: Supplemental Materials for Chapter 2

Biases in Exon Capture Process

We start with an observation that if the exome capture has no bias and the distribution of reads is uniform across the exome, the distribution of depth of coverage (average number of reads per basepair) will appear to follow a Poisson distribution, with mean equal to variance. However, if exon capture biases are present, the variance to mean ratio will inflate above 1, a situation called overdispersion. Thus, we use the variance to mean ratio, called overdispersion factor φ, as a measure of the exon capture bias. The effect of known sources of bias, i.e. GC-content and sequence mapability, is considered and a final justification for using paried-sample comparison approach is given. All of the data presented here are based on the melanoma samples described in the paper and the methods.

Within one sample, there is a substantial amount of biases, and this is reflected through the Poisson Q-Q plot (**Figure S2.1**) and the overdispersion factor of 9.5. Part of this variability comes from the exons with zero coverage. We have examined theses exons with no coverage (available at http://genome.ucla.edu/~fah/ExomeCNV/supplement/SureSelect_No_Coverage_Exons_G3362.bed) and found that they correspond to homologs or repetitive sequences and have low mapability scores (based on ENCODE CRG mapability score). The aligner program discarded reads mapped to these regions because of the ambiguity. Even when removing these zero coverage exons, the amount of bias observed remains substantial (overdispersion factor of 8.9).



A B

Figure S2.1: Poisson Q-Q plot and histogram shows poor fit to Poisson distribution and a high level of overdispersion.

Effect of Overdispersion on Specificity and Sensitivity of ExomeCNV

The effect of overdispersion to Poisson distribution can be modeled through the quasi-likelihood approach in which the variance is allowed to inflate: $\sigma^2 = \varphi\ \mu$. Thus the transformed depth-of-coverage ratio statistic (see Methods for definition and derivation) becomes:

$$t(\rho,\phi) = \frac{\mu_Y R - \mu_X}{\sqrt{\sigma_Y^2 R^2 + \sigma_X^2}} = \frac{\lambda R - \rho\lambda}{\sqrt{\phi\lambda R^2 + \rho\phi\lambda}} = \frac{(R-\rho)}{\sqrt{R^2+\rho}}\sqrt{\frac{\lambda}{\phi}}.$$

That is, the overdispersion factor φ affects the statistic t by directly scaling down the average depth-of-coverage λ by a factor of φ. In other words, overdispersion will reduce the power and accuracy of prediction as though the depth-of-coverage is reduced by the factor φ. For example, an exon with 35x depth-of-coverage and an overdispersion factor of 5 will have the same power of prediction as an exon with 7x coverage but no overdispersion. **Figure S2.2**

illustrates the effect on the ROC curves for prediction of a duplication event on an exon size 500bp with 35x depth-of-coverage.

However, it is noteworthy that this calculation is only true under an assumption that taking the ratio of depth-of-coverage does not get rid of the overdispersion (the biases). We will soon see that this is fortunately not the case, and taking the ratio of depth-of-coverage of the same exon does significantly reduce the overdispersion effect.

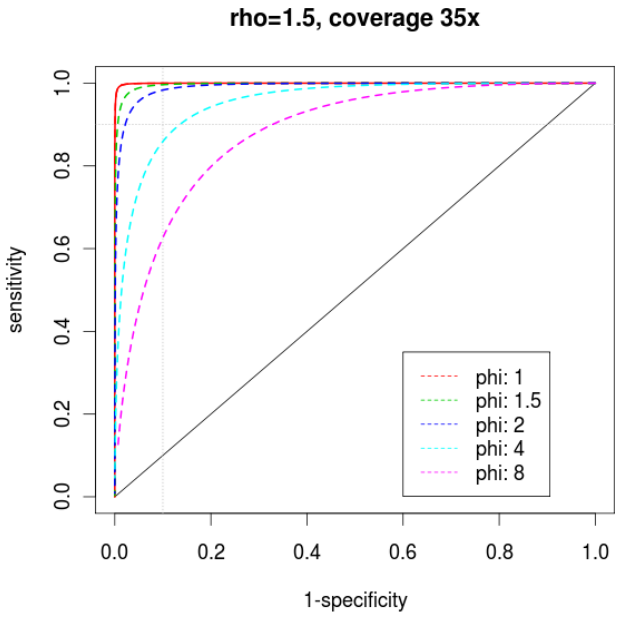


Figure S2.2: Effect of overdispersion on specificity and sensitivity of detecting a duplication event on an exon of size 500bp with 35x depth-of-coverage. The solid red curve is the ROC curve for the case where there is no overdispersion, and the subsequent curves are for the cases with increasing overdispersion factors.

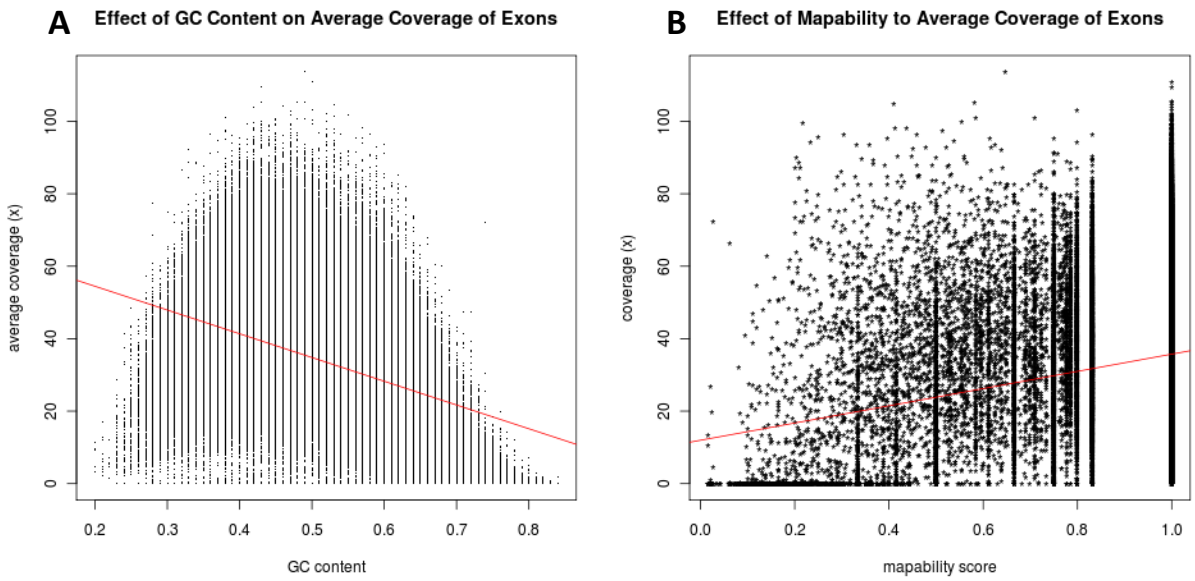


Figure S2.3: Effect of GC-content and mapability on depth-of-coverage of exons.

Effect of GC-Content

Figure S2.3A shows that sequences with extreme GC contents (especially high GC-content) tend to have lower depth-of-coverage. This is in agreement with previous sequencing projects [1,2]. By linear regression, GC-content accounts for 38.52% of the variability in the depth-of-coverage. Correcting for GC-content, we managed to reduce the overdispersion factor to 6.

Effect of Mapability

Mapability, or uniqueness of DNA sequence, can affect the efficiency of sequence alignment. Repetitive sequences have low mapability score and tend to be harder to align, resulting in generally lower depth-of-coverage. On the other hand, high mapability scores indicate unique sequences, and higher depth-of-coverage is observed. Mapability scores (CRG GEM-Alignability of 36mers with no more than 2 mismatches) were retrieved from UCSC Genome Browser, and an average score is calculated for each exon. **Figure S2.3**B shows the

effect of mapability score on depth-of-coverage, with higher scores associated with higher coverage. Ordinary linear regression suggests that mapability accounts for 2.43% of the variance in depth-of-coverage distribution. Correcting for both GC-content and mapability reduces the overdispersion factor to 4.

Taking the Ratio of Depth-of-Coverage Reduces Probe-specific Biases

Although GC-content and mapability can help explain some of the extra variability in depth-of-coverage, more biases remain unexplained. Since our exon capture was done using microarray probes, probe-specific biases such as capture efficiency can have strong effect on the number of mapped reads. Because of the lack of mean to measure and correct for these biases directly, we considered the usefulness of the indirect approach of taking the ratio of depth-of-coverage. Our assumption is that if the effect of biases on depth-of-coverage of an exon is consistent across samples, taking the ratio of depth-of-coverage of the exon from two independent samples will reduce the biases. We looked at the distribution of depth-of-coverage of an exon across 10 exomes available internally and measured the overdispersion factor of the exon. Each overdispersion factor measures how well the depth-of-coverage of the exon follows Poisson distribution, and it turned out that most of the exons have overdispersion factors of less than 1 (mean 0.91, median 0.74, and 3rd quartile 1.17). This implies that it is reasonable to model depth-of-coverage by Poisson distribution and that taking the depth-of-coverage ratio from two samples will reduce the biases.

Justification for the depth-of-coverage ratio threshold used for calling CNV

In calling CNV, we need to identify a cutoff $r(\rho)$ which yields desired minimum specificity and/or sensitivity for testing a particular copy-number ratio ρ at a particular exon with some depth-of-coverage and length. Solving the equations in the main paper Section 2.1 for R, we derive the cutoff value for an α-level test:

$$r_\alpha(1) = \frac{\lambda + t_\alpha\sqrt{2\lambda - t_\alpha^2}}{\lambda - t_\alpha^2},$$

where

$$t_\alpha = \begin{cases} \Phi^{-1}(\alpha) & \text{if } \rho < 1, \\ \Phi^{-1}(1-\alpha) & \text{if } \rho \geq 1 \end{cases}$$

And the cutoff value for a test of power at least $1 - \beta$ is:

$$r_\beta(\rho) = \frac{\rho\lambda + t_\beta\sqrt{\rho(\lambda - t_\beta^2 + \rho\lambda)}}{\lambda - t_\beta^2},$$

where

$$t_\beta = \begin{cases} \Phi^{-1}(1-\beta) & \text{if } \rho < 1, \\ \Phi^{-1}(\beta) & \text{if } \rho \geq 1 \end{cases}$$

If $\rho > 1$, that is we are considering a duplication event, a test rejects when the observed coverage ratio $R > r_{\text{cutoff}}$. Conversely if $\rho < 1$, that is we are considering a deletion event, a test rejects when $R < r_{\text{cutoff}}$.

Figure S2.5 shows a graphical representation of the relationship among α, β, $r_\alpha(1)$, and $r_\beta(\rho)$. From **Figure S2.5**, we note that the exon represented by the red curve does not have sufficient coverage to achieve the desired specificity $(1 - \alpha)$ and sensitivity $(1 - \beta)$ simultaneously, whereas the exon represented by the green curve does have sufficient coverage. If $\rho > 1$, R increases in the direction of the arrows, and the inequality $r_\alpha(1) > r_\beta(\rho)$ indicates

sufficiency of the coverage (as is the case for the green line), and the reverse indicates insufficiency of the coverage (as is the case for the red line). If an exon does not have sufficient coverage, we refrain from declaring CNV for that exon.

In the case when an exon has enough coverage to call CNV, there are multiple possible cutoff values $r \in [r_\beta(\rho), r_\alpha(1)]$, and one can choose to optimize r for sensitivity, specificity, or a function of the two, e.g. area under curve (AUC) = (sensitivity + specificity)/2. We allow the user to choose which option to optimize when performing the test as each option is suitable for different applications.

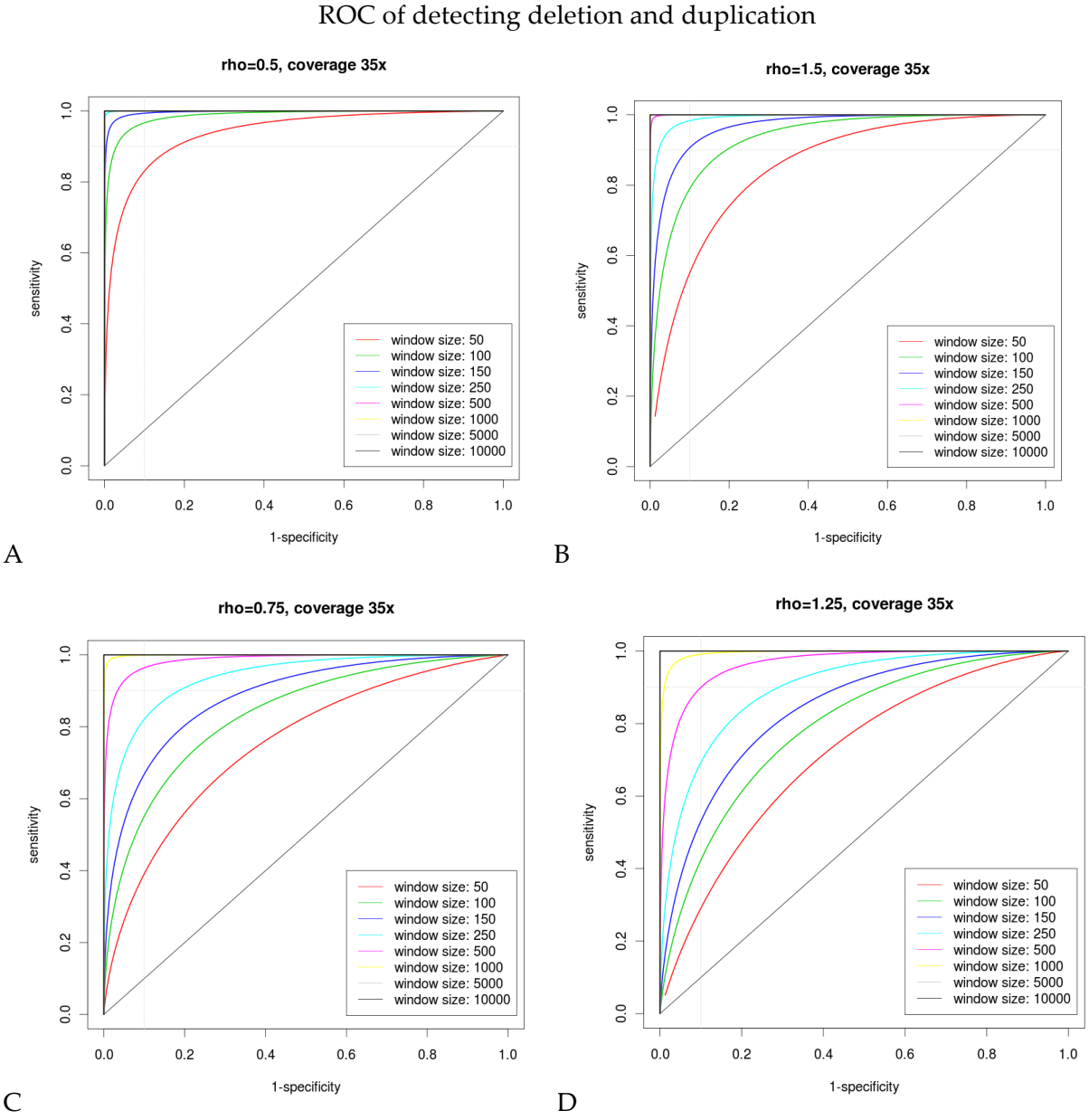


Figure S2.4: ROC curves for detecting (A,C) deletion and (B,D) one copy duplication. The depth-of-coverage is fixed at 35x and read length at 70bp. The dotted gray lines correspond to 95% specificity and sensitivity. Generally, it is more difficult to detect amplification than it is deletion. And at 35x coverage, deletion event (A) can be detected at 95% sensitivity and specificity for exons of size 100 or less while duplication (B) can be detected for exons of size 250 or less. In the presence of admixture of normal genomes, the copy-number ratio ρ tends to one. For example, at the admixture rate of 50%, $\rho = 0.75$ for deletion and $\rho = 1.25$ for duplication. The power and sensitivity decreases as a result of the admixture, and the ROC curves show that (C) deletion can be detected at 95% sensitivity and specificity for exons of size 500bp or less and (D) duplication can be detected for exons of size less than 500bp.

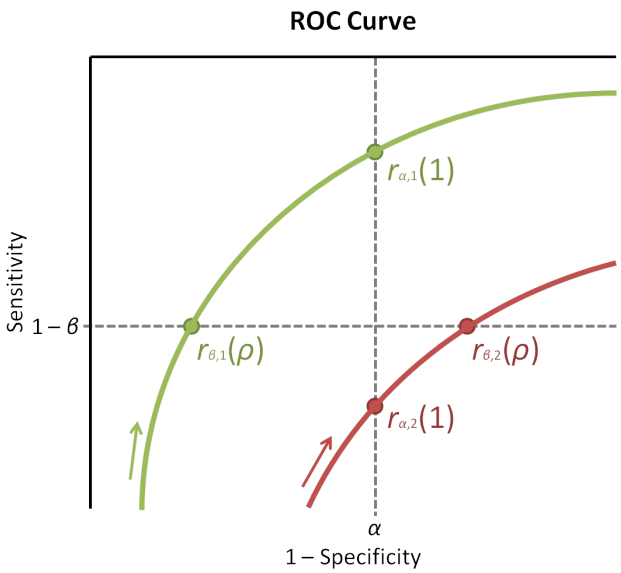


Figure S2.5: A schematic sketch showing relationship among α, β, $r_\alpha(1)$, and $r_\beta(\rho)$. The red curve is an ROC curve of an exon with insufficient coverage to achieve desired specificity and sensitivity $1 - \alpha$ and $1 - \beta$, while the green curve is one with sufficient coverage. The arrows indicate the direction in which the copy-number ratio R increases. Thus, an exon has sufficient coverage to call CNV when $r_\alpha(1) > r_\beta(\rho)$, and not otherwise.

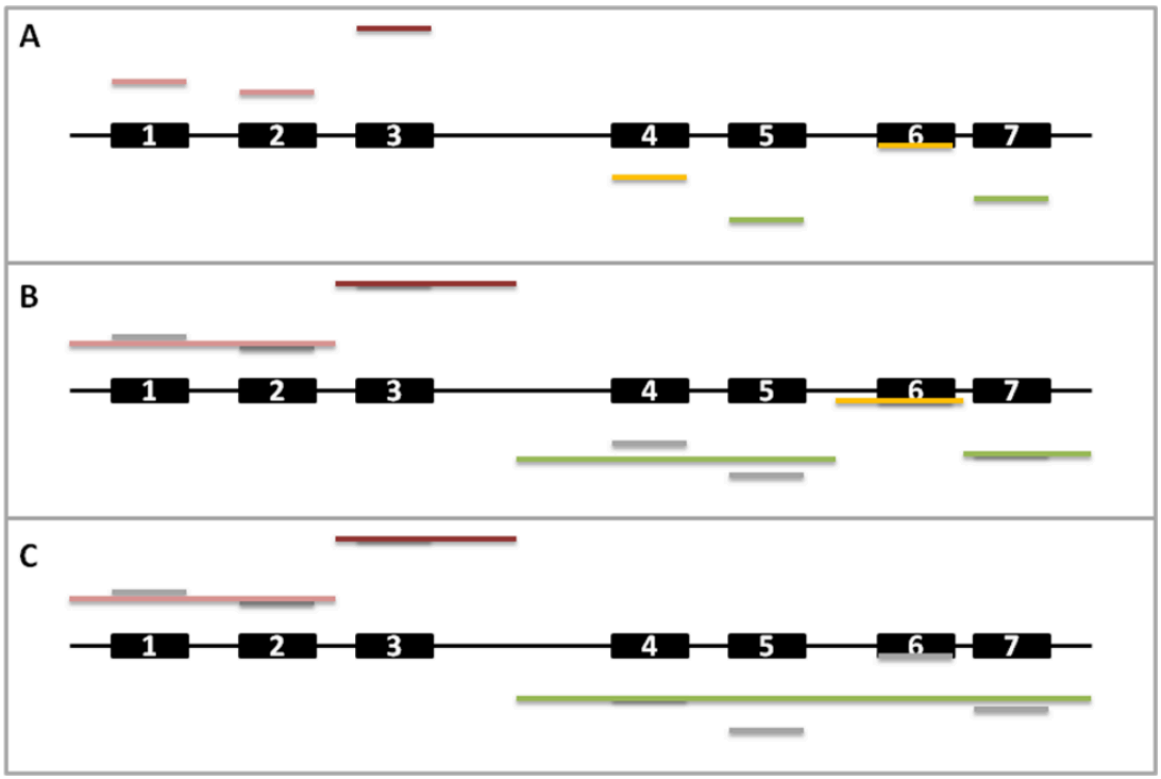


Figure S2.6: Sequential Merging Procedure.

Sequential Merging Procedure

This illustrates three levels of granularity of segmentation. (A) At the finest level of granularity, copy-number variation is assessed at each exon individually. Exons 1 and 2 are called as duplicated (copy number=3) while Exon 3 is called as amplified (copy number > 3). Exons 5 and 7 are called as deleted while Exons 4 and 6 are called as copy-neutral, either because they are truly copy-neutral or because of the lack of power. (B) Circular Binary Segmentation (CBS) subdivides the genome into segments at fine granularity. Exons 1 and 2 merge into one duplicated segment while Exon 4 and 5 merge into a deleted segment. With the merged segment at this level, there is enough power to call deletion on Exon 4. (C) Final segmentation at the coarsest level. Exons 4-7 merge together into a large deleted segment, giving power to call deletion. The final CNV calls consist of three segments as shown.

Other Statistics for Detecting Segmental LOH

Here we discuss the choice of statistics used to detect LOH in a segment produced by the circular binary segmentation (CBS) algorithm. An F-test for equality of variance described in the Methods Section appears to be very sensitive and can detect slight changes in copy-number, which are sometimes caused by non-LOH events such as amplification. Since the increase in variance due to LOH is generally greater than those due to non-LOH events, we can compensate for this over-sensitivity with a conservative p-value threshold. However, since the choice of p-value threshold is quite arbitrary, we considered other statistics: non-parametric Wilcoxon rank-sum test and folded-normal test.

One major challenge in detecting LOH is the fact that non-reference or B-alleles are not always on the same chromosome strand. Thus, the direction of the deviation of B-allele frequencies (BAFs) from 0.5 cannot be used, and the BAFs cannot be combined directly in any meaningful way. The only information we can gain from BAFs is the magnitude of the deviation: $|BAF - 0.5|$ or $|BAF_{case} - BAF_{control}|$ (**Figure S2.8**). For a non-LOH region, we may assume that BAF is normally distributed with mean 0.5 and certain variance σ^2 (**Figure S2.7**A), and so $|BAF - 0.5|$ and $|BAF_{case} - BAF_{control}|$ follows the corresponding folded-normal distribution (**Figure S2.7**C,D). For an LOH region, BAF will have a bimodal distribution centering around 0.5, with each half approximable by a normal distribution (**Figure S2.7**B). These observations serve as basis on which Wilcoxon test and folded-normal test are developed.

The motivation for using Wilcoxon test arises from an observation that the deviations of BAF $|BAF - 0.5|$ in LOH and non-LOH regions follow two distributions with different means (**Figure S2.8**B). Thus we can use Wilcoxon rank-sum test, which is non-parametric and insensitive to model assumption, to detect the difference in the means of the BAF deviation between case and control. The Wilcoxon test appears to be very sensitive, achieving as high as 96.89% sensitivity in detecting LOH in the melanoma sample. However, because of its sensitivity, it also detects changes in the deviation of BAF due to other non-LOH copy-number changes, resulting in low specificity (50-61%). In general, the Wilcoxon test performs very similarly to the F-test.

We attempt to address over-sensitivity problem encountered by the Wilcoxon test and F-test by developing a test that is less sensitive and is able to distinguish between LOH and non-

LOH BAF shifts. Because the absolute difference in BAF $|BAF_{\text{case}} - BAF_{\text{control}}|$ can be assumed to follow a folded-normal distribution, we can use the folded-normal distribution to test for a significant deviation from the null case where BAF_{case} and BAF_{control} are identically distributed. The mathematical details of the test are outlined here:

Assuming BAF_{case} and BAF_{control} follow the normal distributions $N(\mu_{\text{case}}, \sigma^2)$ and $N(\mu_{\text{control}}, \sigma^2)$, respectively, the difference $BAF_{\text{case}} - BAF_{\text{control}}$ follows $N(\mu_{\text{case}} - \mu_{\text{control}}, 2\sigma^2)$, and the absolute difference $|BAF_{\text{case}} - BAF_{\text{control}}|$ follows a folded-normal distribution with mean:

$$E(|BAF_{\text{case}} - BAF_{\text{control}}|) = \sigma\sqrt{2/\pi}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left[1 - 2\Phi\left(\frac{\mu}{\sigma}\right)\right]$$

and variance:

$$\text{Var}(|BAF_{\text{case}} - BAF_{\text{control}}|) = \mu^2 + \sigma^2 + \left\{\sigma\sqrt{2/\pi}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left[1 - 2\Phi\left(\frac{\mu}{\sigma}\right)\right]\right\}^2$$

where $\mu = \mu_{\text{case}} - \mu_{\text{control}}$, and Φ is the standard normal cumulative distribution function (CDF).

When BAF_{case} and BAF_{control} are identically distributed, $\mu = \mu_{\text{case}} - \mu_{\text{control}} = 0$, i.e. a half-normal distribution. Thus, the CDF of the folded-normal becomes:

$$P(X \le x) = 1 - P(X > x) = 1 - 2P(Z > x) = 2\big(1 - P(Z > x)\big) - 1 = 2\Phi(x) - 1$$

and the p-value for the folded-normal test is given:

$$P(X > x) = 1 - P(X \le x) = 2\big(1 - \Phi(x)\big)$$

where x is a realization of the standardized $X = |BAF_{\text{case}} - BAF_{\text{control}}|/\sigma$. In practice, we use $x = \text{average}(|BAF_{case} - BAF_{control}|)/\text{s.e.}(|BAF_{\text{case}} - BAF_{\text{control}}|)$.

As expected the folded-normal test gives more conservative results, achieving 97.52% specificity with 54.27% sensitivity. When combined with the position-wise binomial test (described in the main text), we can improve the sensitivity to 67.55% while lowering specificity to 88.01%.

The choice of test for LOH depends on the application and users' tolerance to false positive and false negative. Finally, we believe that there exists a more efficient test that makes use of other information, such as predicted CNV status and haplotype, and more sophisticated computational techniques, such as Hidden-Markov Model (HMM), but we consider this to be beyond the scope of our present study.

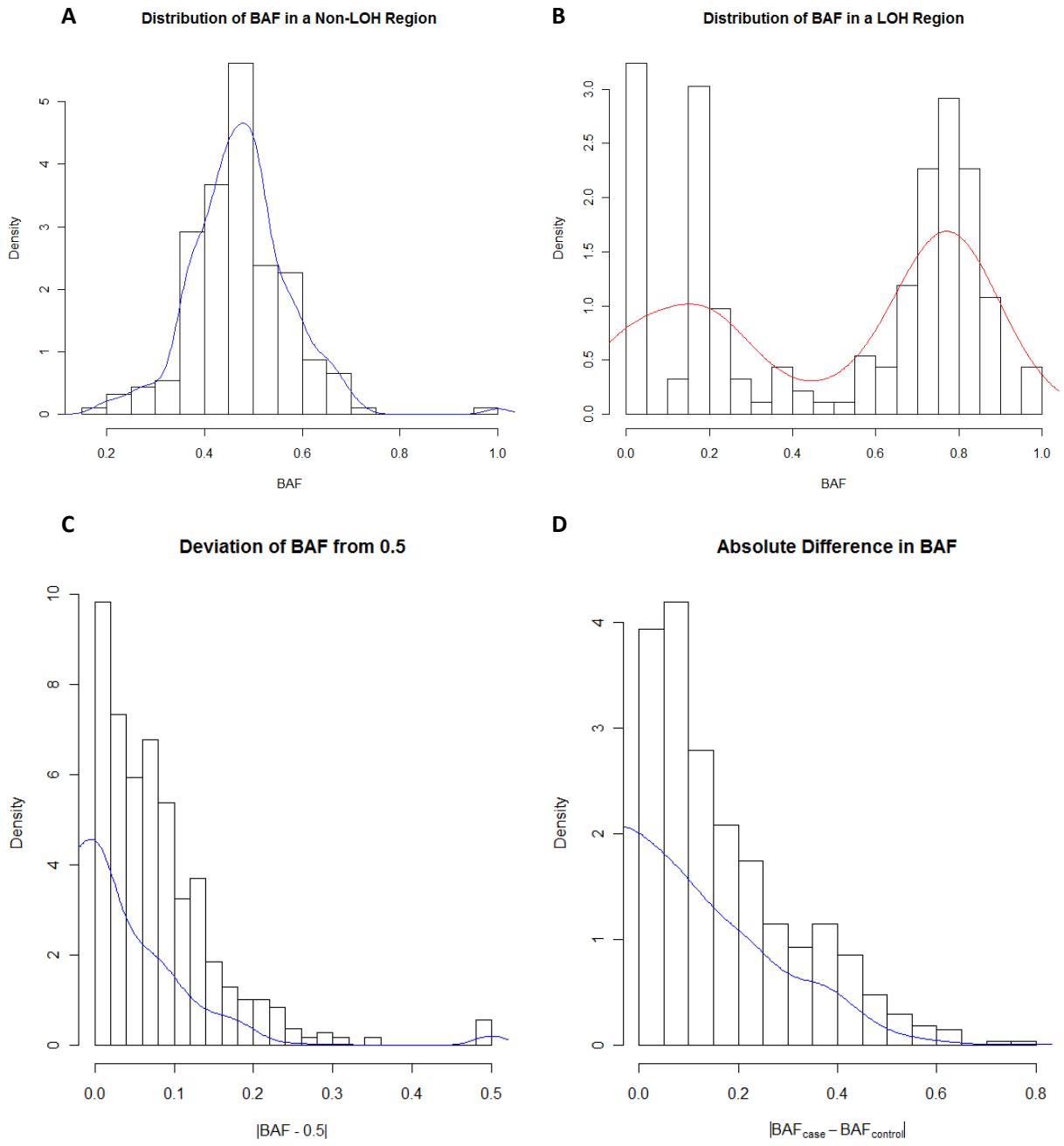


Figure S2.7: Distribution of B-allele Frequencies (BAF) in non-LOH and LOH regions. (A) The distribution of BAF in a non-LOH region follows a normal distribution with mean 0.5, while (B) the distribution of BAF in an LOH region follows a bimodal mixture of normal. (C), (D) The deviation of BAF from 0.5 $|BAF - 0.5|$ and the absolute difference in BAF ($|BAF_{\text{case}} - BAF_{\text{control}}|$) follow folded-normal distributions.

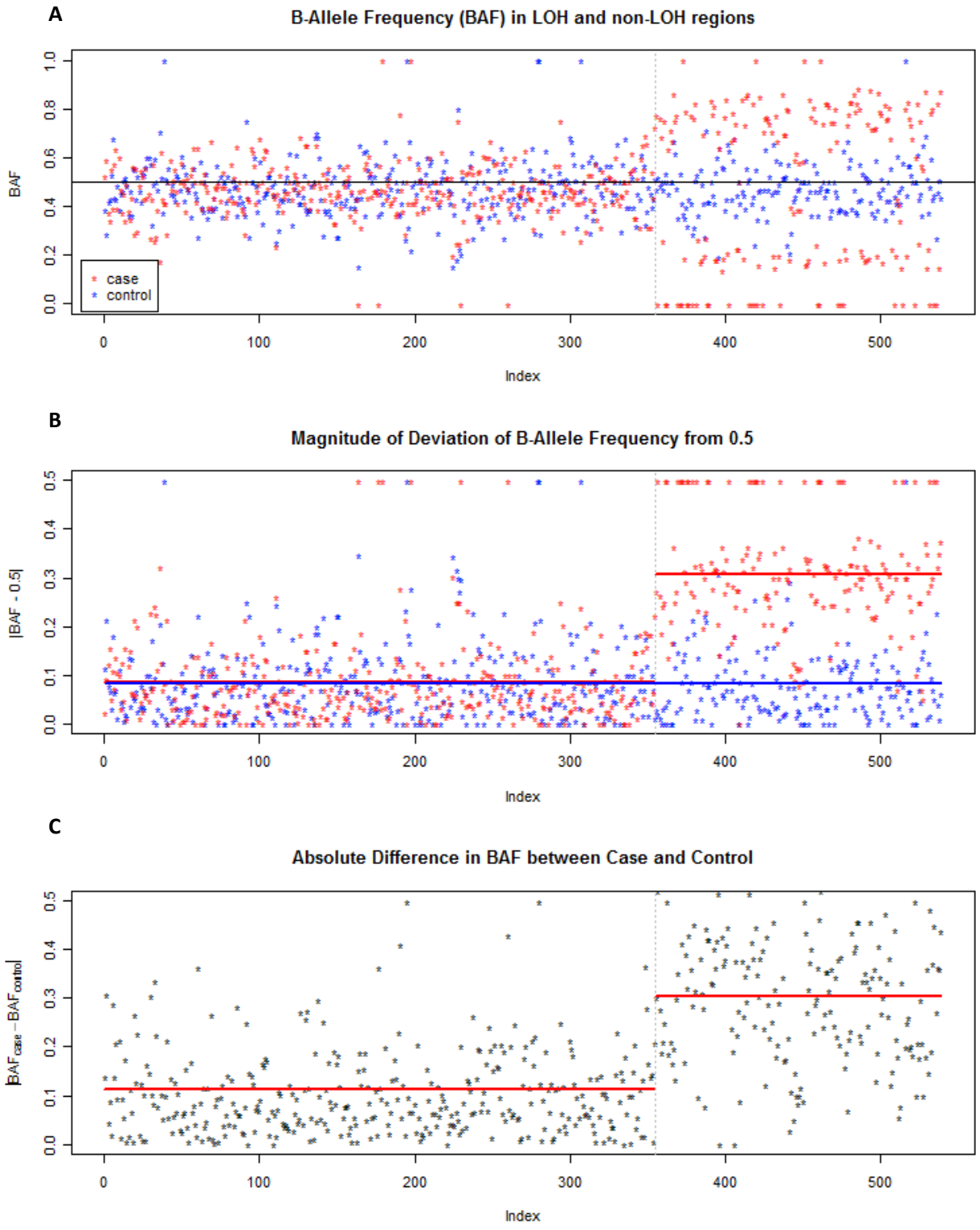


Figure S2.8: BAF, Deviation of BAF from 0.5, and Absolute Difference in BAF between Case and Control.

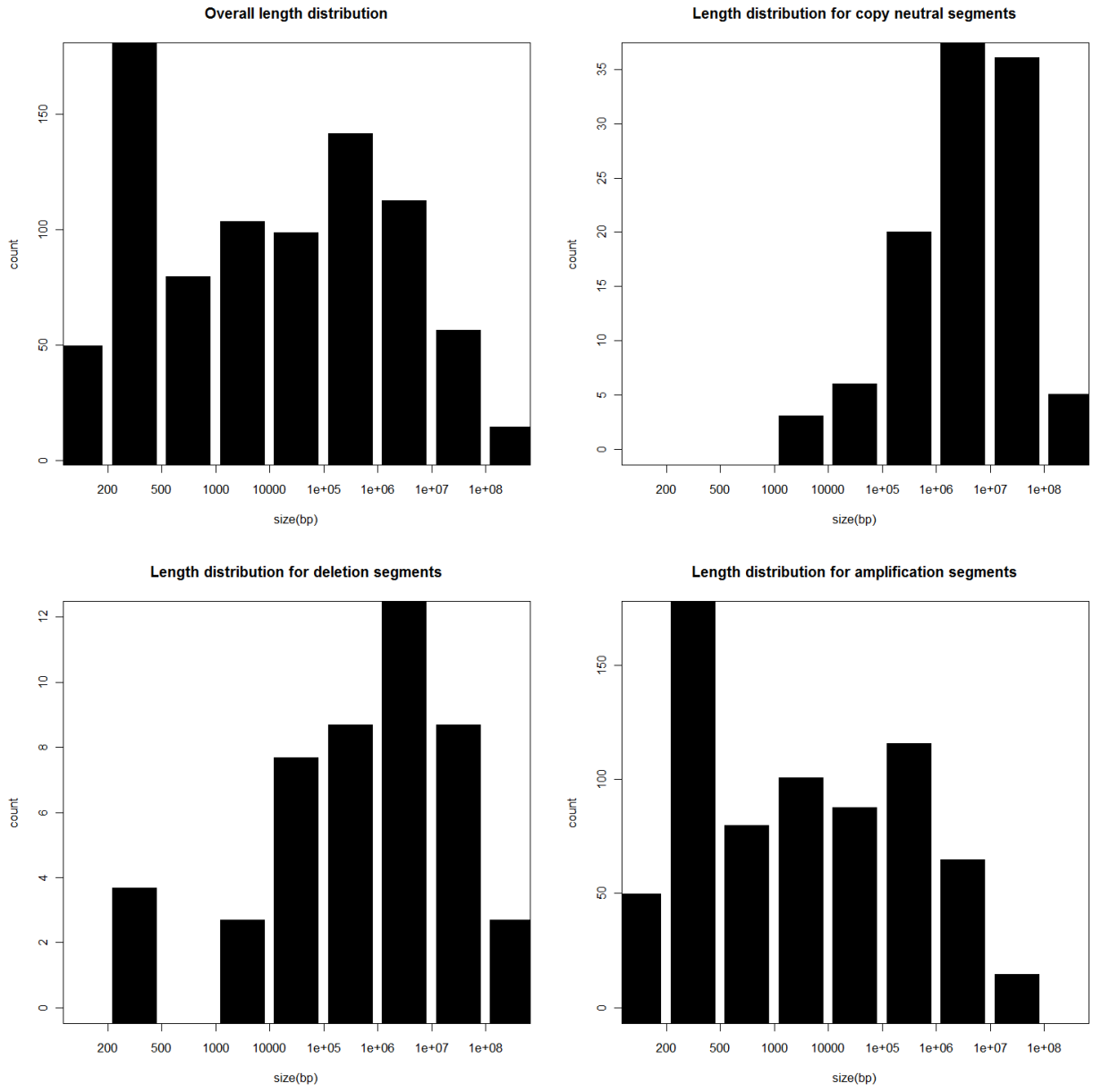


Figure S2.9: Size distribution of CNV calls in the melanoma samples. The sizes of the CNV segments from ExomeCNV range from single exon (120bp) to whole chromosome (chr 10 and 18). The distribution of amplified segments are biased toward smaller segments because ExomeCNV distinguishes segments with evidence for higher copy numbers (e.g. 3, 4, 5) apart while considering all deleted segments the same. Thus all the deleted segments were merged together forming larger segments while amplified segments remain fragmented. We cannot verify the biological validity of this behavior at this point.

In the analysis of melanoma samples, ExomeCNV was allowed to call higher copy numbers (3, 4, 5, etc.). These higher number amplifications were observed in small segments, often single exon, and were not merged together during the CBS-sequential merging step. Thus, the amplified segments have higher number of small segments (< 500bp). In our validation, we treated these higher copy number amplifications as one group.

We processed sequencing data from two lanes of the same matched skin sample run, call these Lane1 and Lane2. The average depth-of-coverage of Lane1 and Lane2 are comparable (21.3x and 20.4x, respectively). Since Lane1 and Lane2 were from the same library and processed exactly the same way, Lane1 should have no copy-number change with respect to Lane2, and any CNV calls are false positives. We ran ExomeCNV, treating Lane1 as case and Lane2 as control, and counted the number of exons falsely classified as deleted or amplified. Here we set estimated admixture rate to 30%. Adjusting minimum specificity from 0.9 to 0.999, we observe empirical specificities as reported in the main text (**Figure S2.10**). When the minimum specificity was set to 0.999, ExomeCNV made 2865 calls; all of which were copy neutral. At minimum specificity of 0.99, ExomeCNV made 5738 calls: 25 deletions, 5711 copy-neutral, and two duplications. And at minimum specificity of 0.9, 27366 calls were made: 1903 deletions, 25073 copy-neutral, and 390 duplications.

Figure S2.10: Lane1-Lane2 Test for False Positive. The plots show log depth-of-coverage ratio and CNV calls (yellow = copy-neutral, red = amplification, green = deletion) by chromosome. Gray dots are the exons with insufficient coverage to be called on their own. (A) is the results from setting minimum specificity to 0.99 and (B) from setting minimum specificity to 0.9. Note that (A) yields higher sensitivity but also lowers the sensitivity by calling CNV for a less number of exons.

Sex-Chromosome Test for False Negative

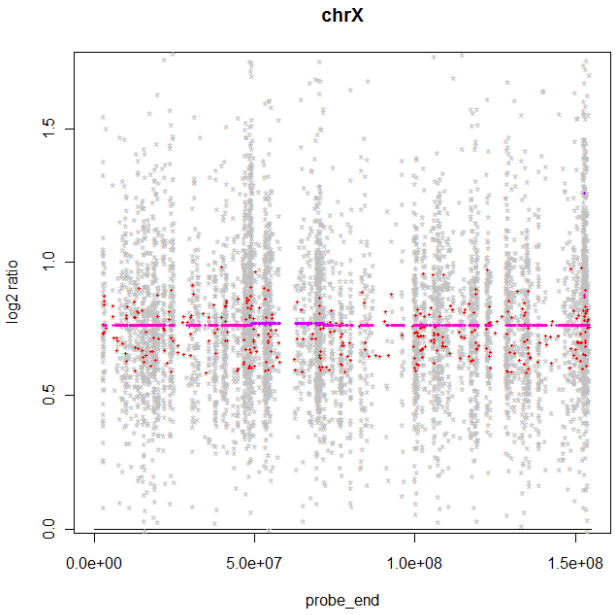


Figure S2.11: Sex-Chromosome Test for False Negative. Exons on Chromosome X are called as amplified by ExomeCNV with most (>99.9%) of the log depth-of-coverage ratio greater than 0. Only Chromosome X is shown here because almost all of the exons in Chromosome Y have depth-of-coverage of zero in male exome and cannot be meaningfully represented (the few exons with non-zero depth-of-coverage have very low depth-of-coverage, are called as deleted by ExomeCNV, and are sequencing errors). Different shades of red represent strength of the signal.

Since we knew the exact copy-number of sex-chromosomes in normal males and females, we used two internally available exomes, one male and one female, to test if ExomeCNV can detect this copy-number difference. The two individual exomes are from individuals with no evidence of sex-liked copy-number aberration. Treating the male exome as case and female as control, ExomeCNV correctly identified Chromosome X as being "amplified" and Chromosome Y as being "deleted" with no false negative (**Figure S2.11**). Here we set minimum specificity to 0.9999 in exon-wise calling and 0.9 in segment-wise calling (which is the default settings) and set admixture rate to 0.

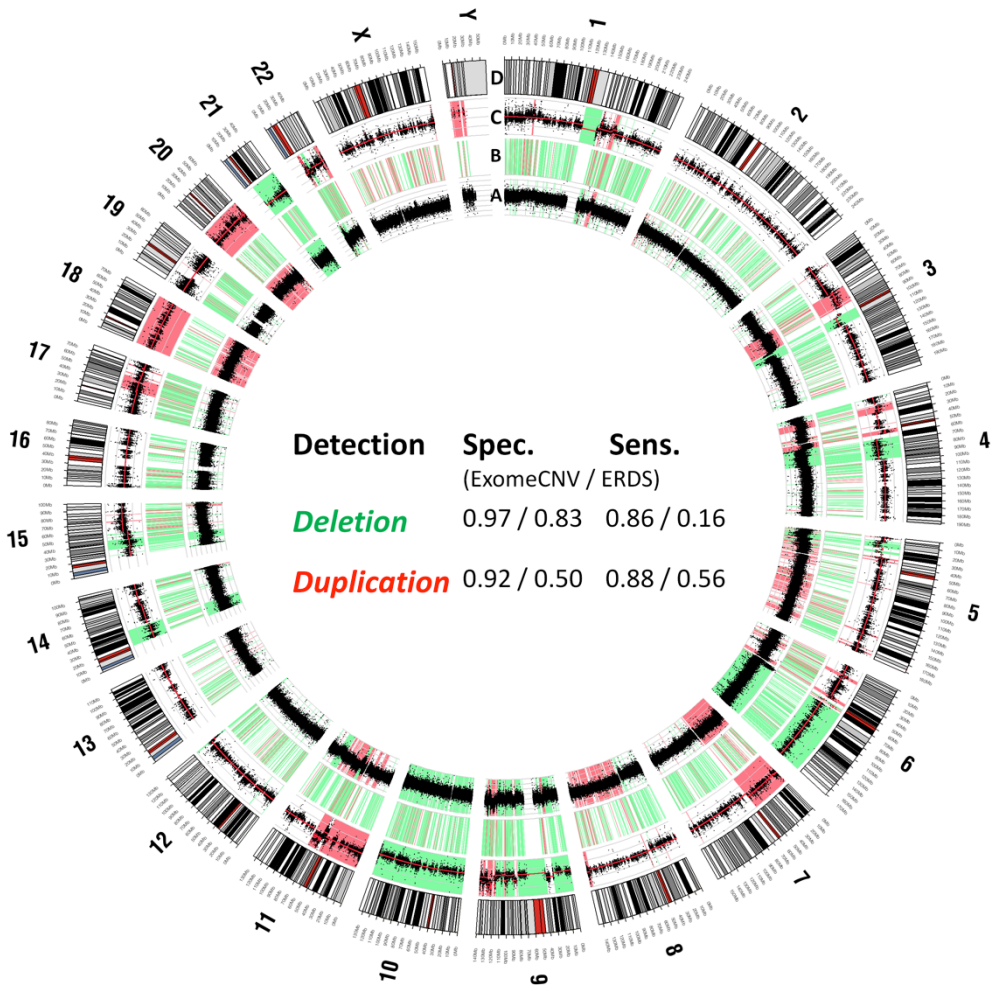


Figure S2.12: Analysis of Melanoma and Paired Normal Samples. Comparison of CNV calls from exome sequencing data using ExomeCNV and ERDS, compared to calls from genotyping array. The most outer ring (D) shows the chromosome ideograms in a pter-qter orientation, clockwise with the centromeres in red. From inside to outside, each data track represents (A) Log R Ratio (LRR) from genotyping array with the region of gain highlighted in red and the region of loss highlighted in green; (B) CNV calls from ERDS, the region of gain highlighted in red and the region of loss highlighted in green (C) log ratio of tumor and normal depth-of-coverage with the segment mean in red line, the region of gain highlighted in red and the region of loss highlighted in green. The CNV for the chromosome Y were not called for the genotyping data as genoCN (the algorithm used to call CNV from Omni-1) is not designed to analyze chromosome Y. The table in the middle summarizes best achievable specificity and sensitivity of ExomeCNV and ERDS in detecting CNV relative to CNV calls from Omni-1 array assessment.

Using Pooled Sample as Control

In many applications, for example in identifying germline CNVs, we do not have matched normal sample to compare the exome of interest with. We propose using a pooled sample as control sample. Samples to be pooled have to be from libraries of the same type (e.g. paired-end, single-end) and processed by the same exon capture and sequencing protocols. This usually means all of the samples should be processed at a particular site. Pooling can be done by averaging depth-of-coverage of each exon across all exomes weighting each exome equally or by their total depth-of-coverage.

We have pooled eight available exomes which were processed and sequenced in the same manner as the melanoma samples used in our study. The depth of coverage on each exon was averaged and used as a "normal" control to compare against normal and tumor exomes from a melanoma patient. The variance of the pooled depth-of-coverage decreased, as expected by the central limit theorem (Figure S2.13). The ExomeCNV results are shown in **Figure S2.14**.

As cautioned in the main text, there is a problem in validating such analysis. First, we cannot ascertain that the pooled exome actually represents an exome with normal copy number of two. These exomes we used were from patients with various genetic abnormalities which may have abnormal copy number variants. Moreover, 4 of the 8 exomes came from the related individuals and may share substantial amount of germline CNVs which would skew the distribution of the pooled copy number. And even if all samples are from unrelated individuals, there might be common CNVs in the population that distorts the "normality" of the pooled sample.

Moreover, a CNV call from this analysis could arise because of many factors, and there is no direct way to validate the results. A CNV, say deletion, found by comparing the tumor exome against pooled exome may arise from 1) somatic deletion in tumor exome 2) germline deletion 3) duplication in pooled exome or 4) false positive. Since we do not have a germline CNV profiling of the subjects, there is no simple way to assess validity of this approach. Thus, we left this as a speculation in our discussion.

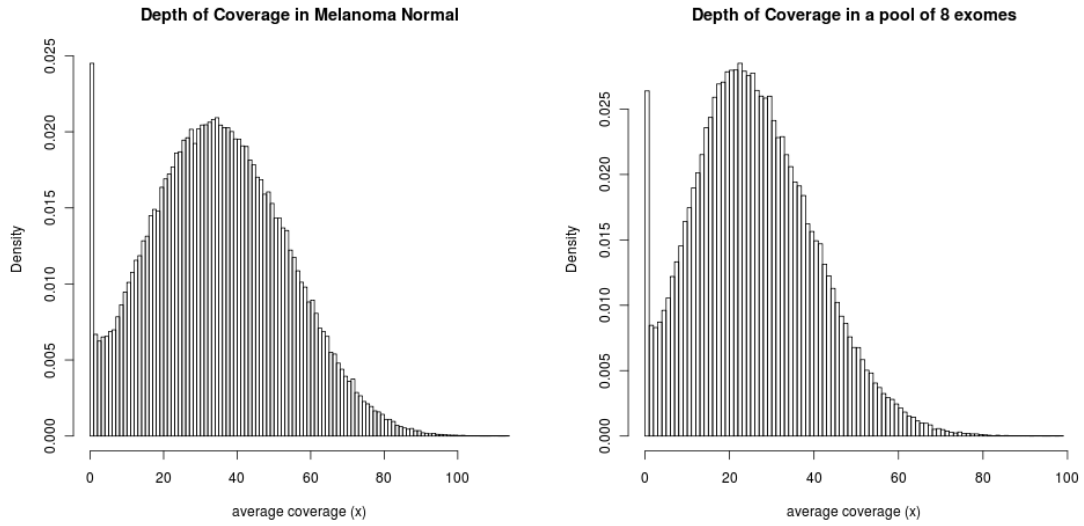
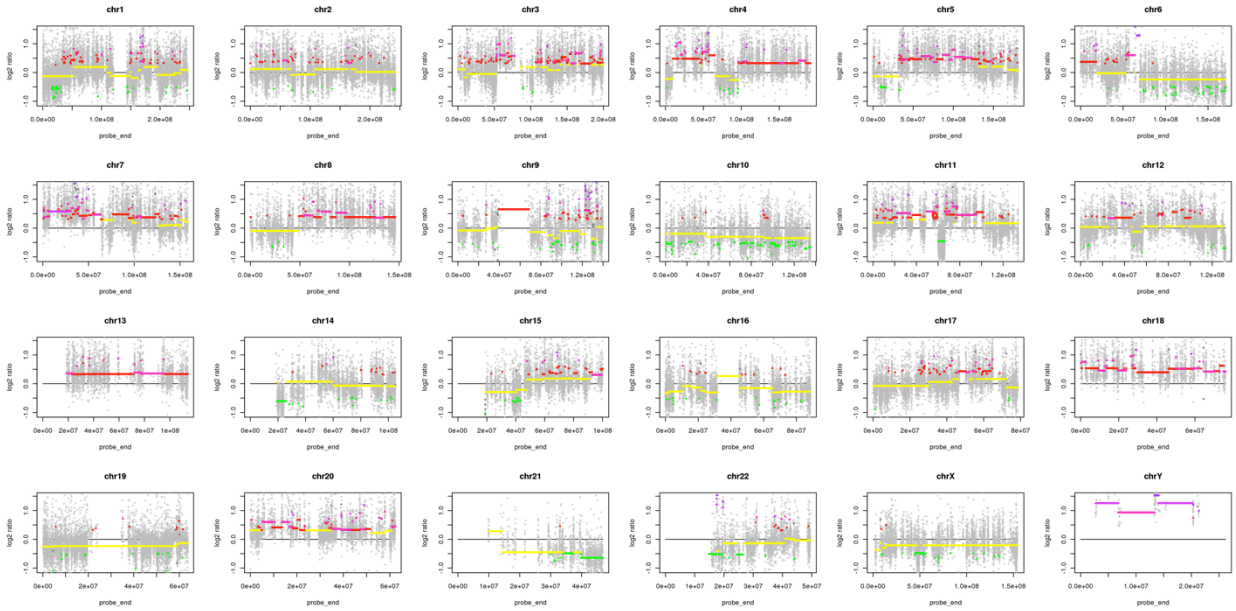


Figure S2.13: Pooling depth-of-coverage of as few as 8 exomes reduces the variance significantly.

A (tumor vs. pooled control)



B (normal vs. pooled control)

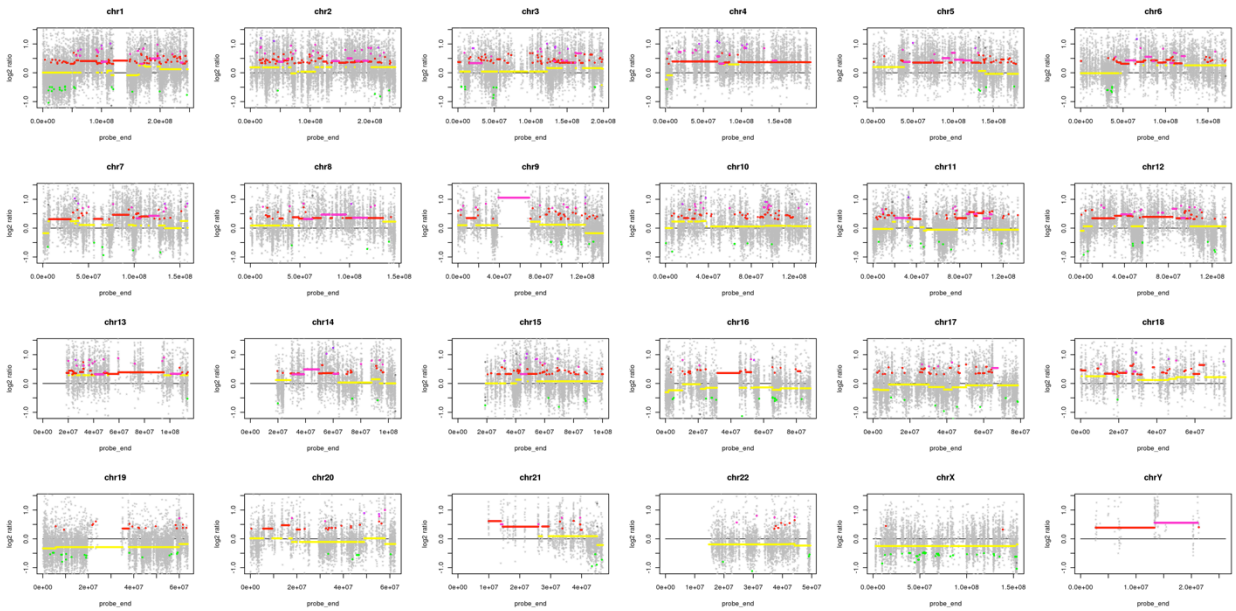
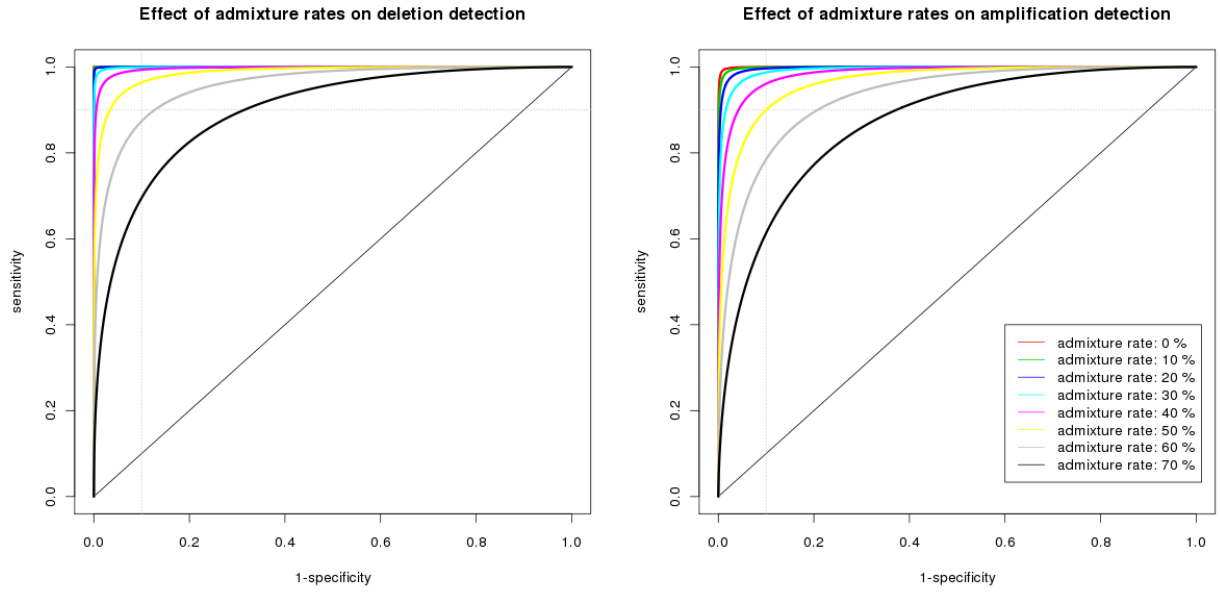


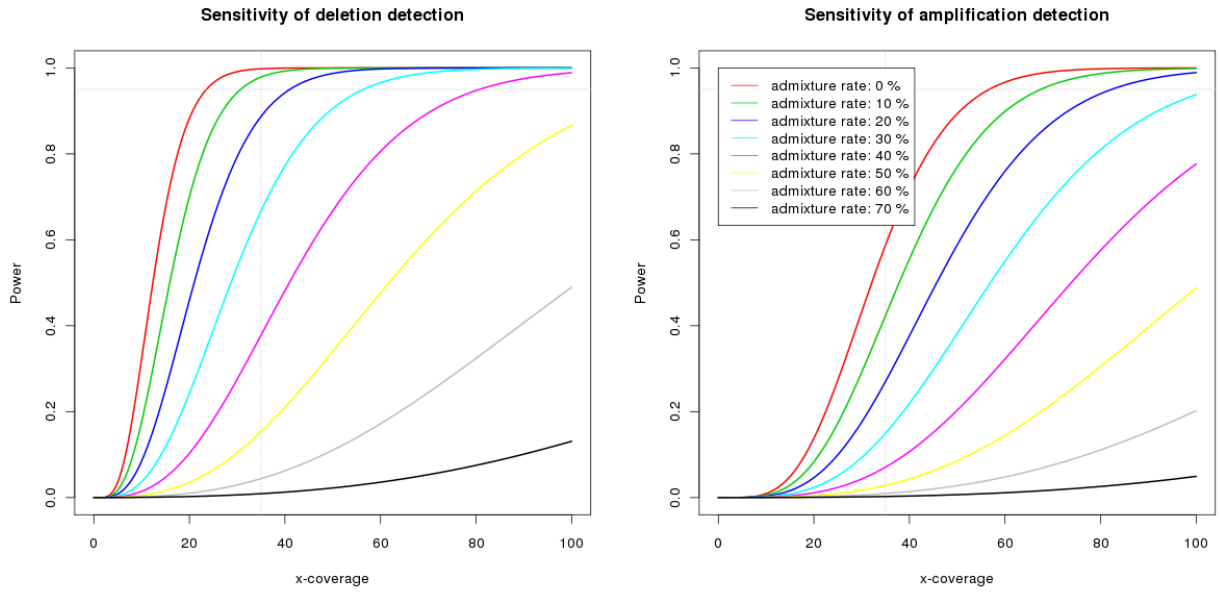
Figure S2.14: ExomeCNV results from comparing melanoma tumor (A) and matched normal (B) with the pooled sample. The amplification observed in chromosome Y is likely due to the imbalance composition of male/female in the pooled sample.

Effect of Admixture Rate on Sensitivity and Specificity of CNV Detection

A



B



C

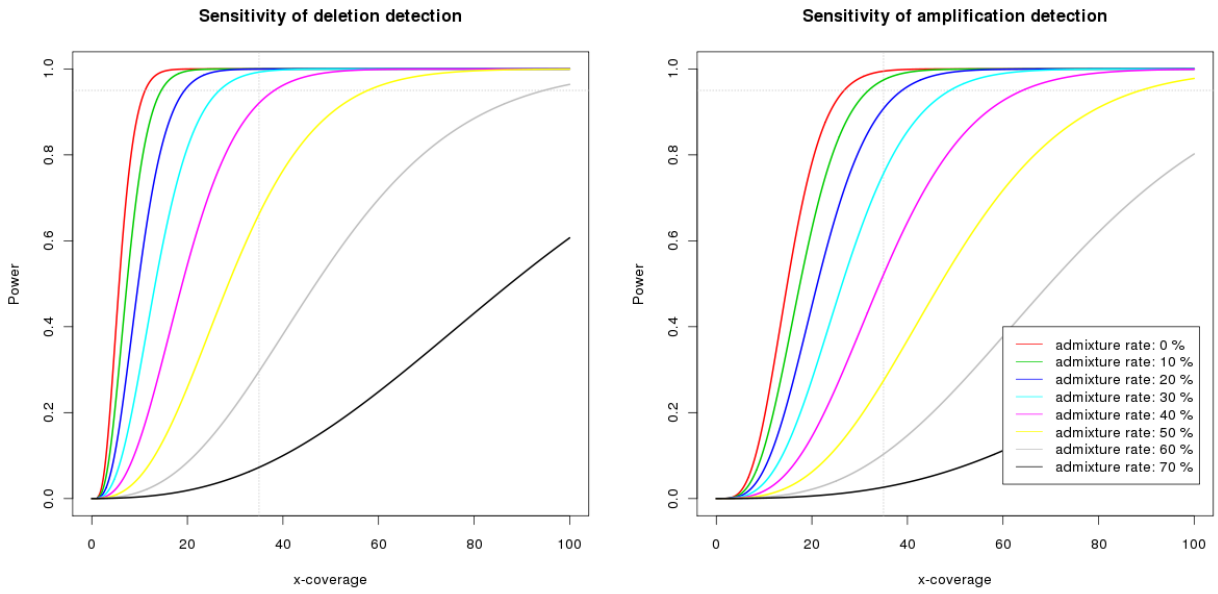


Figure S2.15: ROC and Power curves showing effect of varying admixture rates. (A) ROC curves showing sensitivity and specificity of detecting deletion and duplication of segments size 500bp. (B) Power curves for detecting CNV segments of size 500bp. Power is plotted relative to mean depth-of-coverage in the genomic segment, setting false positive to 1 per genome based on an analytical model of genome-wide power of detection. (C) Same as (B) but for detecting CNV segments of size 1000bp.

Estimation of Admixture Rate from LOH Regions

Because LOH detection method does not require prior knowledge of admixture rate, we can use LOH detection to estimate admixture rate. In particular, because we know that B-allele frequency (BAF) in a LOH region is either $0.5(1 - c)$ or $0.5(1 + c)$ where c is the admixture rate, the value $|BAF - 0.5| = 0.5\ c$. Thus, c can be estimated by

$$\hat{c} = 2\ \text{Average}(|BAF_{\text{LOH}} - 0.5|)$$

where BAF_{LOH} is the B-allele frequency of a LOH region. In our melanoma sample, the admixture rate is estimated to be about 30% by this method, which is in agreement with an estimate from the SNP genotyping arrays.

Appendix 2A Bibliography

1. Campbell, P.J., *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729 (2008).

2. Chiang, D.Y., *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, 99-103 (2009).

Appendix 2B: Supplemental Analysis for Chapter 2

To demonstrate the shortcomings of array-based CNV calling algorithms, we used a state-of-the-art array-based algorithm, genoCN [1], to call CNV from the exome sequencing data. In particular, we are examining the impact of exome sequencing data violating two assumptions made by genoCN: the high density of genotype markers and the distribution of log ratio of the signal intensity. We started by converting exome sequencing data to an appropriate format and applying genoCN.

Method

The objective of this experiment is to use an appropriate array-based CNV calling method to call CNV on the exome sequencing data. Since our "gold standard" CNV calls are from using the state-of-the-art method genoCN on the Illumina Omni-1 Quad BeadChip SNP array data, we converted the exome sequencing data into 59,857 SNPs based on the Omni-1 Quad array annotation. For each SNP, the B-allele frequency (BAF) is the fraction of the total sequencing reads with B allele (minor allele), and the log ratio (LRR) is the log2 of the normalized read counts between tumor and normal samples. Default values were used for other genoCN parameters.

Comparison between the genoCN calls based on the exome sequencing data and the gold standard calls proceeded the same way as other comparisons in the ExomeCNV paper. Namely, the CNV regions were first map to exons with at least three genotype markers (total

3,516 exons), then the true/false positive/negative were calculated in terms of the number of exons correctly/incorrectly identified as copy-number gain (>2)/loss (<2)/neutral (2).

Results

In our melanoma sample, used in the ExomeCNV paper (Sathirapongsasuti et al. 2011), genoCN called 3,103 as loss and 237 exons as gain CNVs based on exome sequencing, while the gold standard (based on the appropriate application of genoCN to the dense SNP array data of the same sample) called 377 exons as gain and 426 as loss CNVs. The total number of exons in consideration is 3,516. These correspond to specificity of 0.955 and sensitivity of 0.252 for amplification and specificity of 0.132 and sensitivity of 0.988 for deletion (Table SA2.1).

	True Positive	False Positive	True Negative	False Negative	Specificity	Sensitivity
Deletion	421	2682	408	5	0.132	0.988
Amplification	95	142	2997	282	0.955	0.252

Table SA2.1: An array-based approach genoCN as applied to exome sequencing data

GenoCN appears to call much fewer amplifications than deletions. This is not surprising as the power analysis of CNV detection suggests lower statistical power to detect amplification than deletion. Because of the lower number of positive calls, the sensitivity is low (0.252) but the specificity is considerably higher (0.955). On the other hand, a much higher number of exons (1,589) are called as deleted, yielding markedly low specificity (0.132) but achieving a considerable sensitivity (0.988).

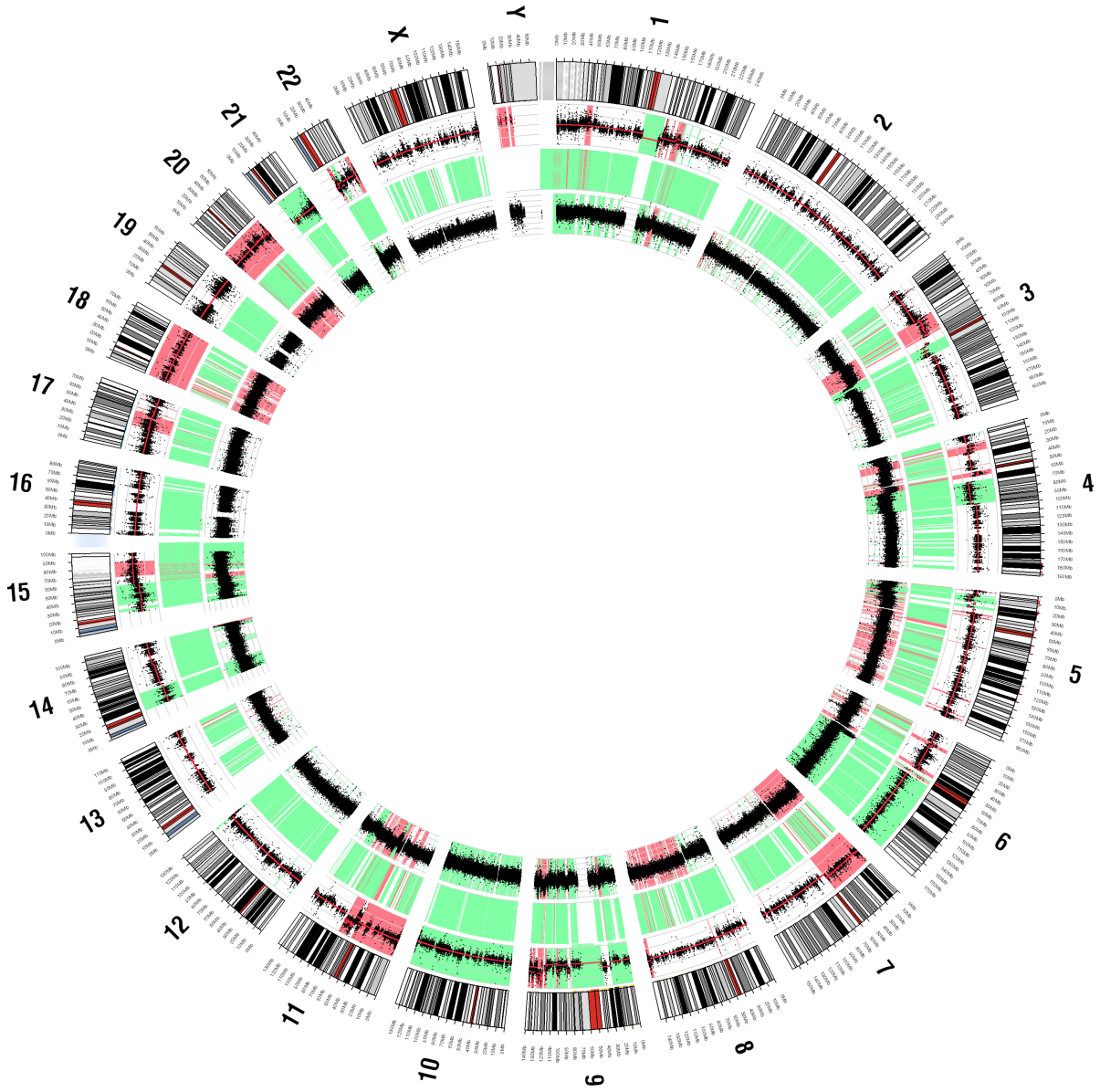


Figure SA2.1: Circos plot comparing CNV calls from three methods: (from outside in) 1) ExomeCNV 2) exome sequencing-based genoCN and 3) SNP array-based genoCN. Using the appropriate application of SNP array to genoCN as the gold standard, ExomeCNV achieves high sensitivity (0.86 deletion, 0.88 amplification) and specificity (0.97 deletion, 0.92 amplification), while the application of exome sequencing data to genoCN achieves low sensitivity (0.25 amplification) or low specificity (0.132 deletion).

Discussion

GenoCN makes two primary assumptions about the input log ratio intensity (LRR) values: the LRR follow a mixture of uniform and normal distributions, and there is at most one state transition between two adjacent SNP probes. The first assumption is not appropriate in the setting of exome sequencing as we and others [2] have shown that the read counts from exome sequencing follows a Poisson distribution. Consequently the log ratio of the read count is not normally distributed, as shown by the QQ plot of the LRR from exome sequencing data from the melanoma samples used in the main analysis (Figure SA2.2).

The second assumption is that there is at most one state transition between two adjacent SNP probes. This assumption is appropriate to make in the setting of the SNP array because of the high-density of genotype probes [1]. In exome sequencing, this assumption is not appropriate to make because there are regions in the genome with few exons (gene deserts). Moreover, exons tend to be more highly conserved and have lower density of SNPs. In our particular dataset, out of 1,140,419 SNPs in the Omni-1 Quad annotation set, only 59,857 SNPs are in the exome (19-fold reduction of information). And some of the regions that were falsely called by genoCN are indeed SNP-poor regions. For example, the regions false called as copy-gain CNVs have on average 4 genotype probes per region, compared to the genome-wide average of 7 probes per region.

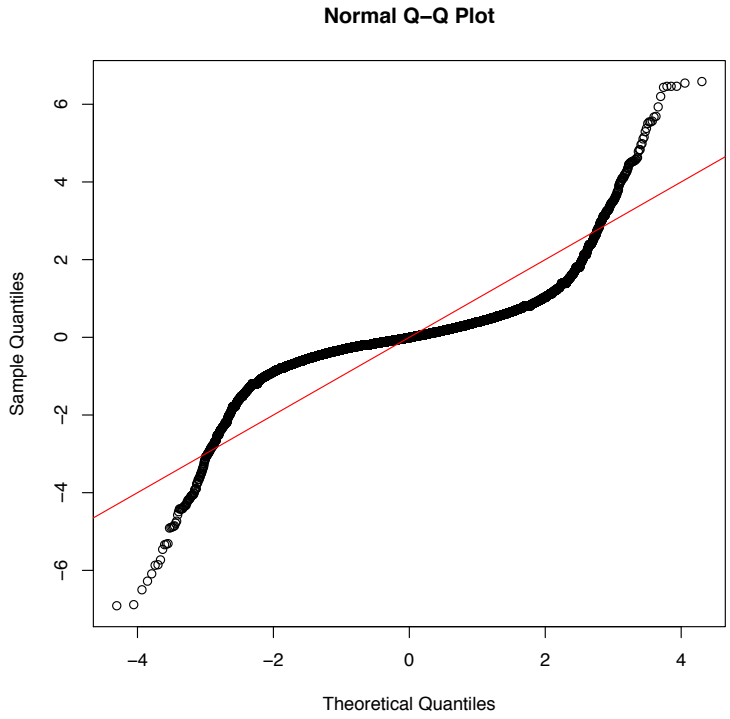


Figure SA2.2: QQ plot of LRR values from exome sequencing. The distribution of exome sequencing read counts has been show to follow a Poisson distribution, and at > 30x, the ratio of the read counts can be approximated by a Cauchy distribution. Thus the log ratio of read counts is not normally distributed, as illustrated by this QQ plot.

The difference between SNP array and exome sequencing can also be seen at the individual probe level. Because of the 19-fold reduction in the number of genotype probes, genoCN is much less reliable. As with any data set, a single measurement by a genotype probe can be unreliable. A comparison of LRR values between the two data types (Figure SA2.3) shows that although there is a strong correlation between the two LRR (Spearman's 0.294), there are clear differences at individual probe LRR. In particular, there appear to be a number of probes with near zero LRR in one but high LRR in the other. This could be because of the normalization of the read counts and probe intensity that regress the values toward the means, making the normal and tumor samples look more alike. While array-based approach can

benefit from borrowing information across probes, with 19-time less number of probes, exome sequencing-based approach can be misleading.

Two illustrative examples highlight the need for a new method that properly model the data are the deleted q-arm of Chromosome 6 and the duplicated p-arm of Chromosome 7. Figure SA2.4 shows the B-allele frequency (BAF) and log ratio intensity (LRR) that were used by genoCN. Clearly the shift of LRR below zero in Chromosome 6 and above zero in Chromosome 7 were evident, suggesting that with a proper modeling of the BAF and LRR, CNVs can be called from the exome sequencing data. As discussed above, the assumptions made by genoCN do not hold in this data, justifying the need for a new method.

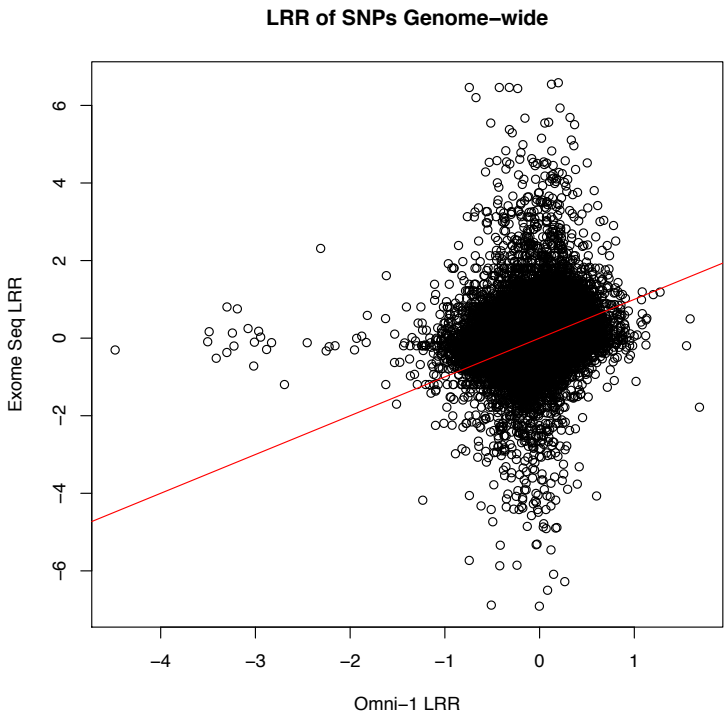


Figure SA2.3: Comparison of log ratio between exome sequencing read counts and SNP array intensity. Despite the overall correlation (Spearman's 0.294), there are a number of probes with discordant LRR. Notably a number of probes have close to zero LRR in one but high LRR in the other.

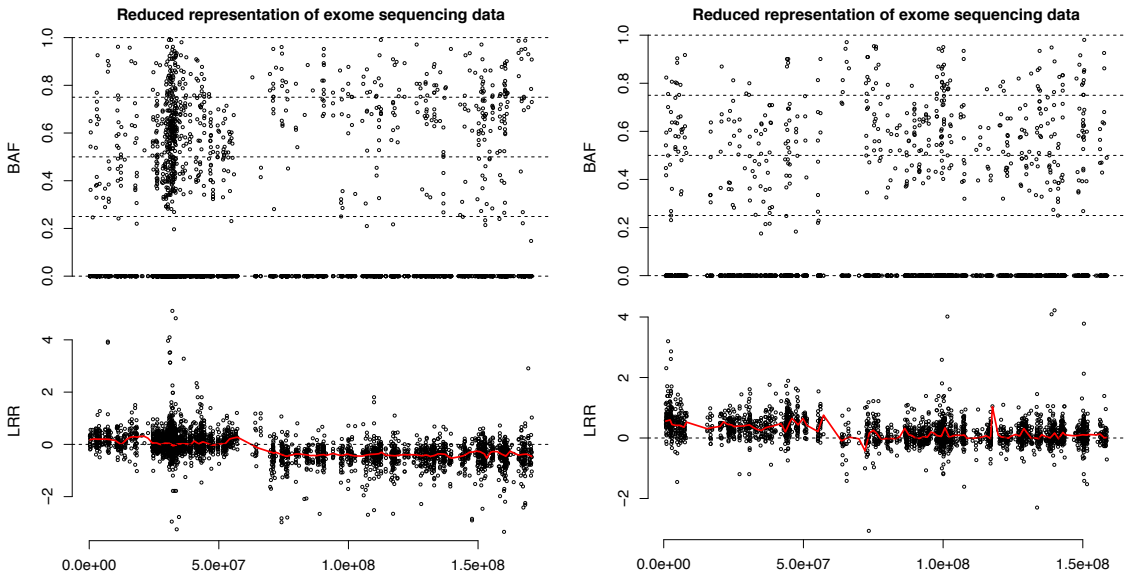


Figure SA2.4: LRR and BAF as input for genoCN. The deletion of the q-arm of Chromosome 6 and duplication of the p-arm of Chromosome 7 are evident in the LRR plot by the shift of LRR below and over the zero line. GenoCN failed to detect both of these events (see also Figure SA2.1). This illustrates that the information about CNV existed in the exome sequencing data but a proper modeling was needed.

Appendix 2B Bibliography

1. Sun W, Sun W, Wright FA, Tang Z, Nordgard SH, Van Loo P, Yu T, Kristensen VN, Perou CM. Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res*. 2009 Sep;37(16):5365-77.

2. Guo Y, Long J, He J, Li CI, Cai Q, Shu XO, Zheng W, Li C. Exome sequencing generates high quality data in non-target regions. ***BMC Genomics*** 2012, **13**:194

Appendix 3A: Supplemental Methods for Chapter 3

Accounting for Compositional Structure in Relative Abundance Data

Data summarized as relative, rather than absolute, abundances such as 16S profiling is known to be challenging to analyze due to compositional measurements. Specifically, since relative abundances are normalized to sum to one (or equivalently rarefied to an arbitrary constant total), an increase in one relative abundance must be accompanied by a compositional decrease in another. This is true even for raw sequence read counts due to the fixed depth of sequencing, which samples proportionally from a total underlying population. Assessing correlation between compositional data by traditional measures such as Pearson's correlation can lead to spurious correlations [1]. Thus, we developed a novel methodology that mitigates the compositional effect in data while assessing significance of an association. In particular, we compare the distribution of correlations from bootstrap sampled data, which represents the confidence interval of the observed correlation, with the null distribution of correlations from renormalized permuted data, which represents the correlation structure arising purely from the compositionality.

In general, assessing significance of an association (e.g. correlation between two vectors) involves comparing an observed value to an appropriate null distribution. Permutation, which breaks any true association in the data, is a common approach for constructing the null distribution. Here, we show that simple permutation does not give an appropriate null distribution for compositional data, as it breaks compositionality and provides anti-

conservative estimates. However, appropriate compositional structure can be re-introduced to permuted data by renormalization, and a null distribution can be obtained from the resulting permutation-renormalization scheme. This approach can be further improved by using bootstrapped confidence intervals for the observed value as opposed to the point observation itself, since the bootstrap process scales the variance of the test according to the severity of compositional effect as manifested through the signal-to-noise ratio.

Compositionality leads to spurious correlation and loss of information

As discussed previously [1], relative abundance data which sum to a constant (e.g. one) can exhibit spurious correlation. For example, in **Figure SM3.1**, the absolute (A and B) and relative (C) abundances of four microbes (b1-4) are shown. Microbes b1 and b2 are uncorrelated in **Figure SM3.1**A, but the relative abundance (**Figure SM3.1**C) shows negative correlation between them because of the compositional effect introduced through normalization. It is then important to account for the compositional effect in assessing correlation between abundance data.

Unfortunately, some true correlation may not be possible to recover. In the example shown in **Figure SM3.1**, although the relationship between b1 and b2 is different in A and B, the two different absolute abundance datasets produce exactly the same relative abundance profiles, masking the difference in the true correlation structure. Thus, in some cases, it may be impossible to uncover the true underlying correlation given the relative abundance data alone. Hence, it is important to stress that because only the relative abundances were observed in our HMP 16S dataset, some true correlation between microbes may not be possible to elicit. And in

evaluating the performance of our method, we need to bear in mind that our aim is to mitigate the effect of compositionality and lessen the confidence in correlations that are likely results of the compositional structure rather than to completely recover all true correlations.

Standard permutation tests result in false positive correlations

One standard procedure to evaluate the significance of a correlation is a permutation test. Unfortunately, the permutation test cannot distinguish spurious compositional correlations, as the permutation process removes all compositional effects and generates a highly anti-conservative null distribution. As demonstrated in an example in **Figure SM3.2**, the permutation test declares all pairwise correlation in a synthetic dataset to be highly significant. This is because the location and scale of the permutation null distribution does not reflect the compositional structure. The distribution always centers at zero with constant standard deviation. A desired null distribution should vary the location to reflect the portion of eventual correlation attributable to compositionality alone and vary the scale to reflect measurement scale based on variation in the absolute abundance.

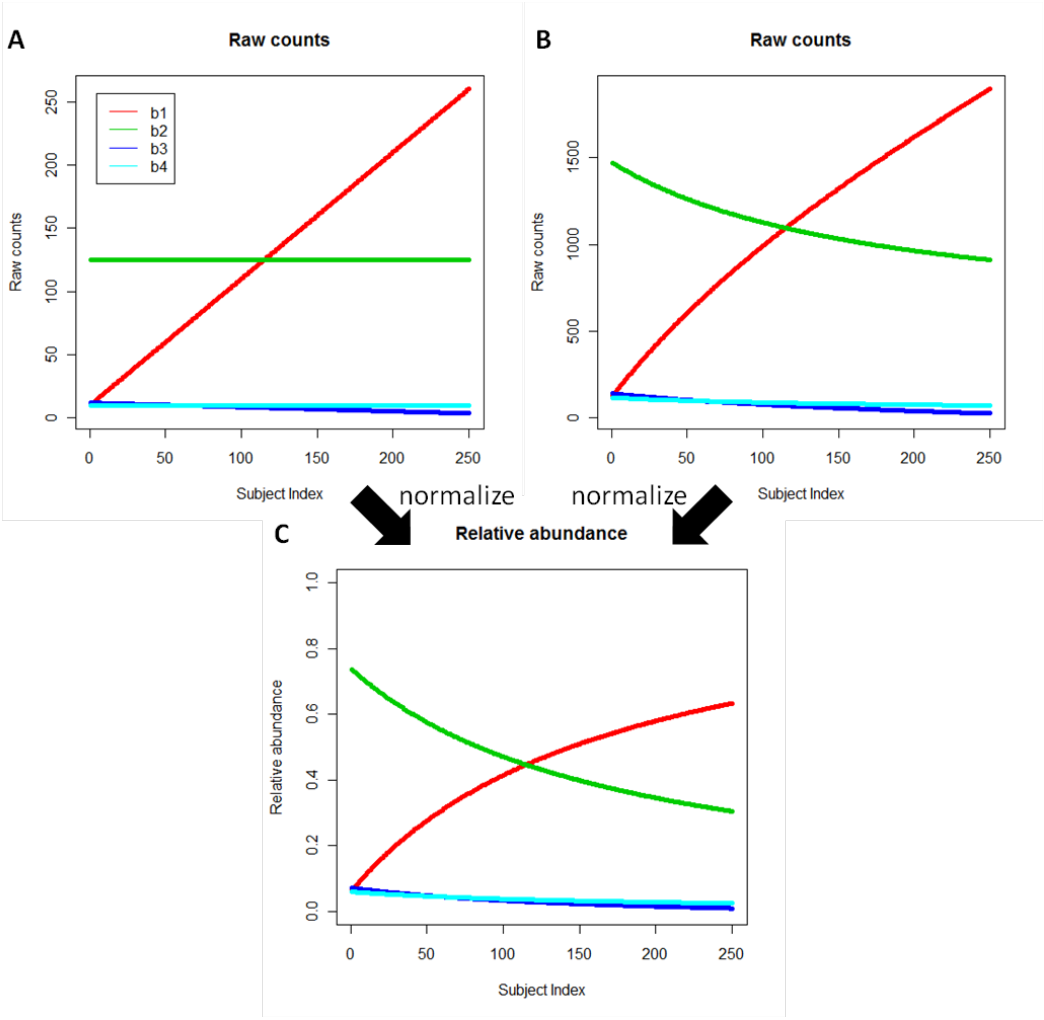


Figure SM3.1: Compositionality induces spurious correlation and causes loss of information. Consider two possible absolute abundance profiles from collections of microbial communities including taxa b1 through b4. 250 synthetic samples (A) and (B) represent absolute abundances (read counts) that, upon normalization into relative abundance (C), produce identical values from which it is impossivle to recover the original information. Hence it is important to note that in even the best-handled relative abundance data, there may be some correlation that cannot completely be recovered.

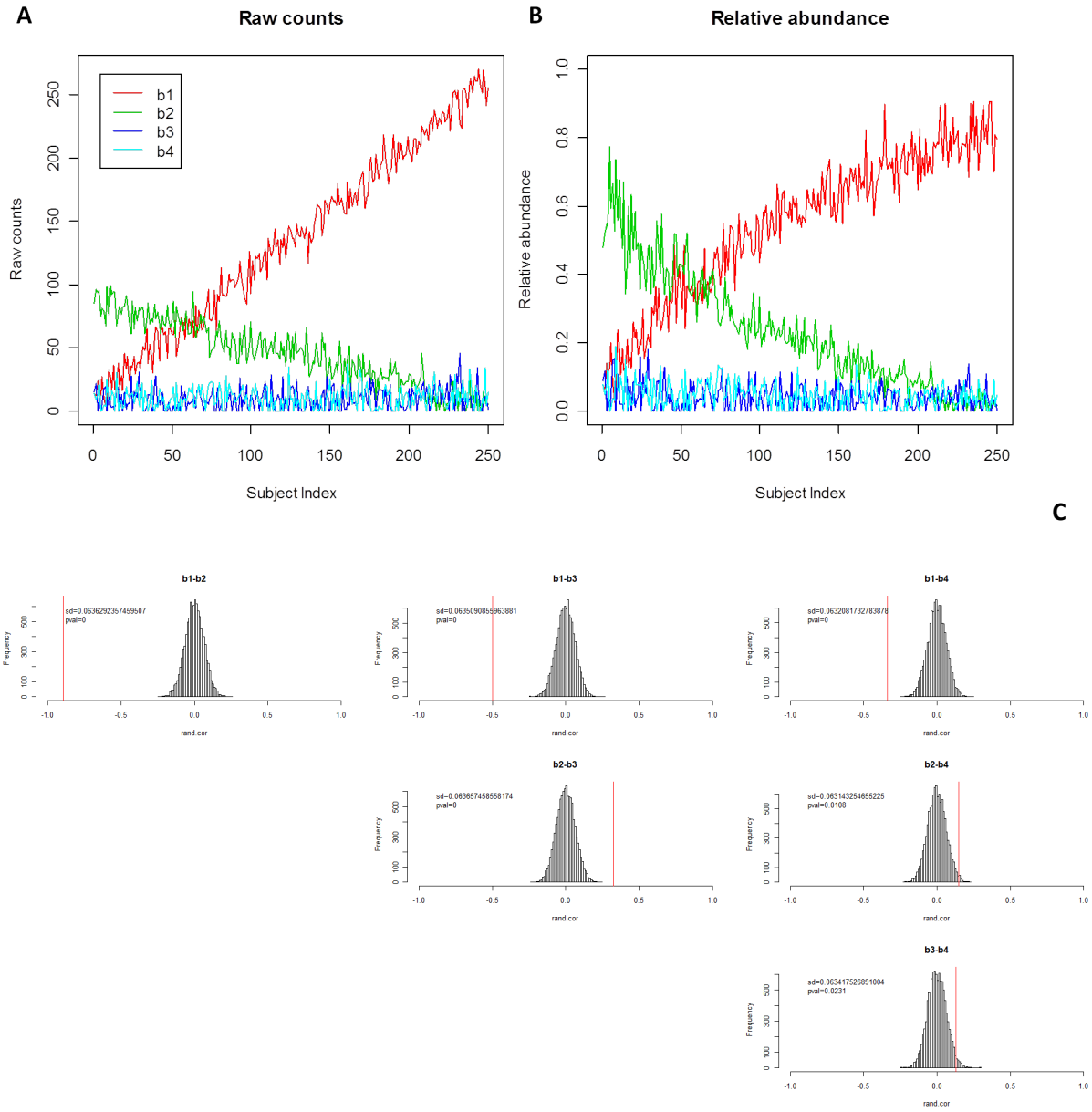


Figure SM3.2: Permutation testing alone cannot distinguish between true and spurious correlations in compositional data. (A-B) Simulated absolute and relative microbial abundances. (C) Pairwise correlations (vertical red lines) based on the relative abundance in (B) for each pair of synthetic microbes. These are accompanied by permutation-based null distributions, standard deviations, and z-test p-values, all of which are highly (and incorrectly) significant due to compositional effects.

The Permutation-Renormalization and Bootstrap (ReBoot) Method

We propose ReBoot: Permutation-Renormalization and Bootstrap Method, a procedure to construct a null distribution that reflects the compositional effect in the abundance data. As shown in **Figure SM3.3**, the method starts with the relative abundance data and consists of two steps: constructing a compositionality-aware null distribution and comparing this with a bootstrap confidence interval around the observed correlation. These steps proceed as follows:

Constructing the compositional null distribution

For each pair of microbes:

1. Permute the relative abundance of the microbes
2. Renormalize the permuted abundance by summing over each sample and dividing the abundance in that sample by the sample sum
3. Calculate the correlation between the renormalized permuted relative abundances of the two microbes
4. Repeat (1)-(3) N times to obtain the compositional null distribution

Constructing the bootstrap confidence interval

For each pair of microbes:

1. Sample with replacement the sample indices and construct a bootstrap resampled dataset
2. Calculate the correlation between the bootstrapped relative abundances of the two microbes
3. Repeat (1)-(2) B times to obtain the bootstrap distribution, which represents the confidence interval around the observed correlation

Finally, compare the compositional null distribution and the bootstrap distribution by z-test with the variance pooled from both distributions using an equally weighted unbiased least square estimate.

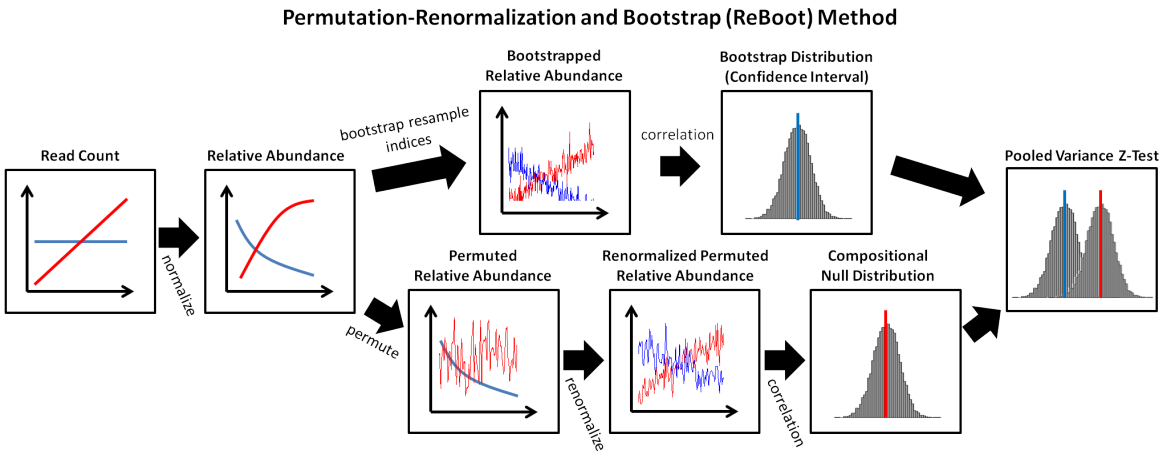


Figure SM3.3: The Permutation-Renormalization and Bootstrap (ReBoot) Method. From the relative abundance data, we construct (1) bootstrap confidence interval around the observed correlation and (2) the null distribution that represent the correlation due to compositionality alone. Contrasting the two distributions through z-test with pooled variance, an appropriate significance level of the observed correlation can be assessed.

Intuition

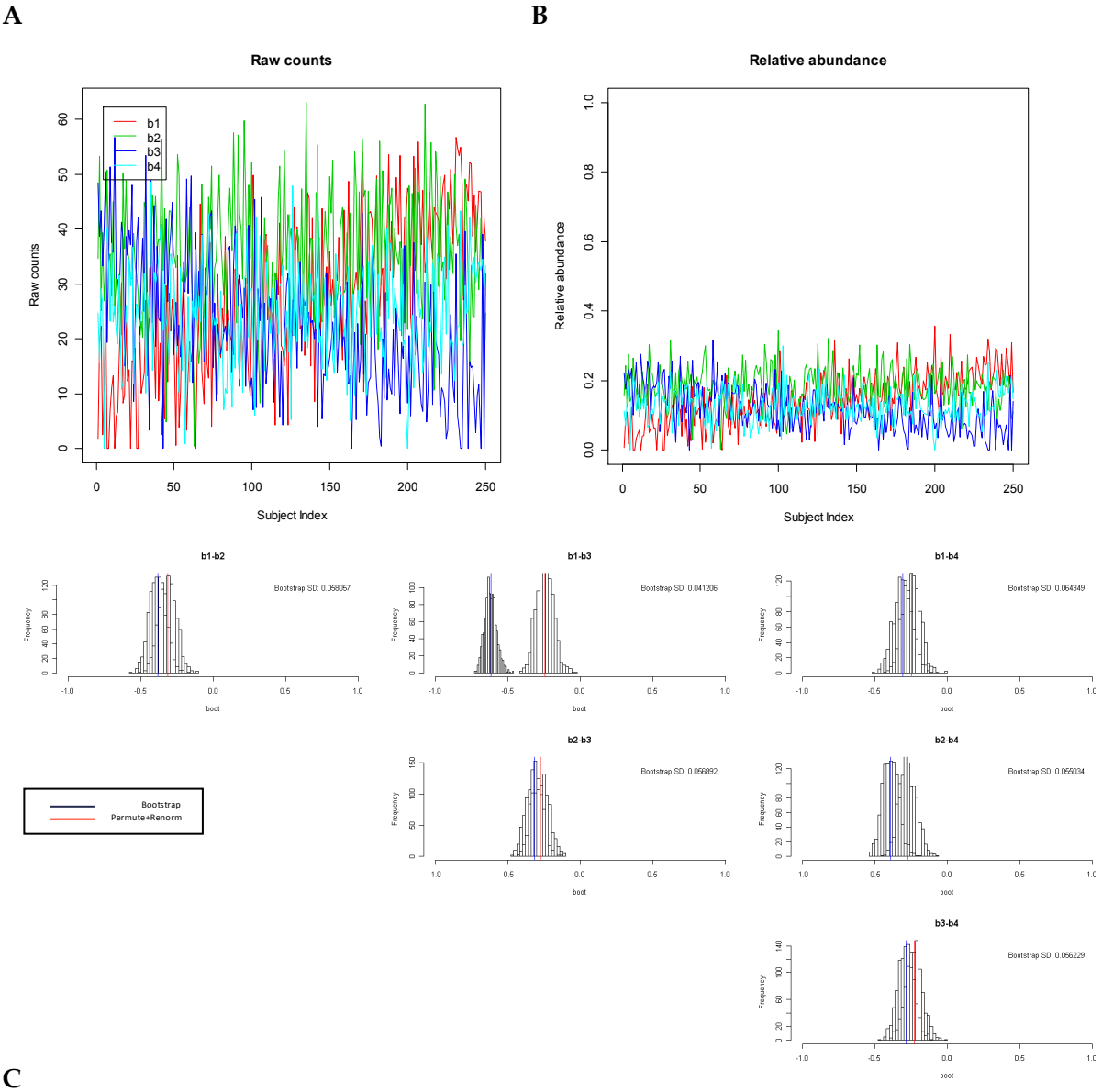
The bootstrap distribution represents the confidence interval around the observed correlation, which will be wider for less abundant organisms due to lower signal-to-noise ratios. The permutation-renormalization null distribution represents correlation due solely to the compositional effect (see the detailed explanation below). The significance of associations over and above those expected from compositionality alone can thus be evaluated by comparing these two distributions. The null hypothesis to be tested is that the two distributions have the same mean, making z-test an appropriate comparison. The variance pooled with equal weight from both distributions provides an unbiased least squares estimate.

ReBoot method works well on simulated data

We demonstrated the performance of the ReBoot Method by assessing significance of the correlation in a simulated dataset (described below). The results are shown in Table SM3.1,

where pairwise correlations and p-values assessed through various methods were compared. In the simulation data (**Figure SM3.4**A), b1 increased linearly while b2 remained constant in all samples; b3 also decreased linearly while b4 remained constant at a lower abundance; 16 other low-abundance microbes were also present at a constant abundance. The all microbes had low abundances and were affected strongly by the noise. The correlation and the permutation test p-values estimated from the absolute abundances were the expected values and were treated as the gold standard against which we compared our results.

Pairwise Pearson correlations calculated from the relative abundances of b1-b2 and b2-b4 pairs differed substantially from the true correlations (Table SM3.1) due to the effects of compositionality as expected. The p-values estimated by a simple permutation test on the relative abundance data overstated the significance of the b1-b2 and b2-b4 correlations but maintained the true significance of the b1-b3 pair. Using the permutation-renormalization null distribution (without the bootstrap distribution) improved the p-value estimates of b1-b2, but the significance level of b2-b4 correlation was still overstated. Finally, when the bootstrap distribution was introduced, the p-values of b1-b2 and b2-b4 were no longer significant, matching the expectation and the synthetic data's gold standard.



C

Figure SM3.4: Combining a renormalized permuted null distribution with a bootstrap confidence interval adjusts the significance of associations between features appropriately according to the degree of compositionality in the data. A synthetic dataset of four simulated microbes was generated as absolute abundances (A) including a single true negative correlation between b1-b3 and uncorrelated taxa b2 and b4. (B) After normalization to relative abundances, b2 and b4 exhibit spurious correlation mitigated by (C) application of the ReBoot procedure to all pairwise comparisons. Blue and red vertical lines represent means of bootstrapped (confidence interval of true association) and permuted (null distribution of association due to compositionality) distributions, respectively. A z-test with pooled variance was used to compare the means of the two distributions, among which only the b1-b3 retains significance (see Table SM3.1), matching the gold standard.

	Pearson correlation		P-value			
	Absolute abundance (true corr.)	Relative abundance (composition)	Permutation test on the absolute abundance	Permutation test on the relative abundance	Permutation-Renormalization	Permutation-Renormalization and Bootstrap (ReBoot)
b1-b2	*-0.00043*	*-0.11471*	*0.490*	*0.039*	*0.067*	*0.114*
b1-b3	*-0.37481*	*-0.45163*	*0.000*	*0.000*	*6.69E-14*	*1.34E-13*
b1-b4	0.016205	-0.04366	0.403	0.247	0.203	0.167
b2-b3	0.022527	-0.03726	0.359	0.292	0.264	0.238
b2-b4	*-0.0603*	*-0.13541*	*0.170*	*0.013*	*0.031*	*0.082*
b3-b4	0.002101	-0.05086	0.518	0.225	0.196	0.170

Table SM3.1: Comparison of ReBoot p-values to a synthetic gold standard

Simulated data

We performed a number of simulations to demonstrate the effect of compositional measurement and assess the performance of the ReBoot method. In the simulations, absolute abundances of 10 to 100 microbes were simulated by adding normally distributed noise to a linear trend. A minimum of 10 microbes were required to capture a realistic compositional effect, while too many simulated microbes will diminish the effect. For each microbe, 250 absolute abundances (the same number as the HMP subjects) were produced as a linear function of the subject index (1-250) with random noise. The standard deviation of the noise distribution is fixed within each simulation. The true correlations between simulated microbes were determined by the correlation of the noise-free linear trends. These absolute abundances were converted to relative abundances by sample-wise normalization.

Renormalization mitigates the compositional effect

As discussed above, an appropriate null distribution should represent the amount of correlation due to compositional measurement alone. While permutation breaks the correlation

structure between pairs of microbes, it also eliminates compositional correlation. During the ReBoot procedure, since only the microbes of interest are permuted, the compositional effect is retained in the other microbes, we can thus reintroduce the compositional structure to the permuted vectors by renormalizing the data so that each sample sum is one. The correlation structure remaining in the renormalized permuted data is the correlation due to compositionality alone.

To illustrate that permutation-renormalization recovers the compositional effect, we performed a simulation (Figure SM5) in which high-abundance microbe b1 is increasing with the sample index, microbe b2 remains constant, low-abundance microbe b3 is decreasing, and the other 97 microbes are constant (and thus uncorrelated) at low abundance. The absolute abundance is normalized into relative abundance whose negative compositional (spurious) correlation between b1 and b2 is clearly visible. For each pair of microbes (e.g. b2-b3 and b1-b2 shown in Figure SM5), permutation was performed to break their correlation, then sample-wise renormalization was performed. The correlations between microbes of interest in the renormalized permuted data were the correlation structure due to compositionality alone and were visible when plotted with index ranked by abundance of the microbes of interest (e.g. b2 and b1, Figure SM5).

It is important to note that renormalization can only recapture the compositional effect in as much as one exists in the relative abundance data. Some compositional effects cannot be removed because of the information lost during the initial normalization, and in these cases ReBoot errs on the side of potential false positives, assuming that correlations greater than

would be expected by chance in relative abundances were also significant in the underlying hidden absolute abundances. Again, **Figure SM3.1** illustrates such a case where it is impossible to fully distinguish compositional, spurious correlation from the original absolute correlations.

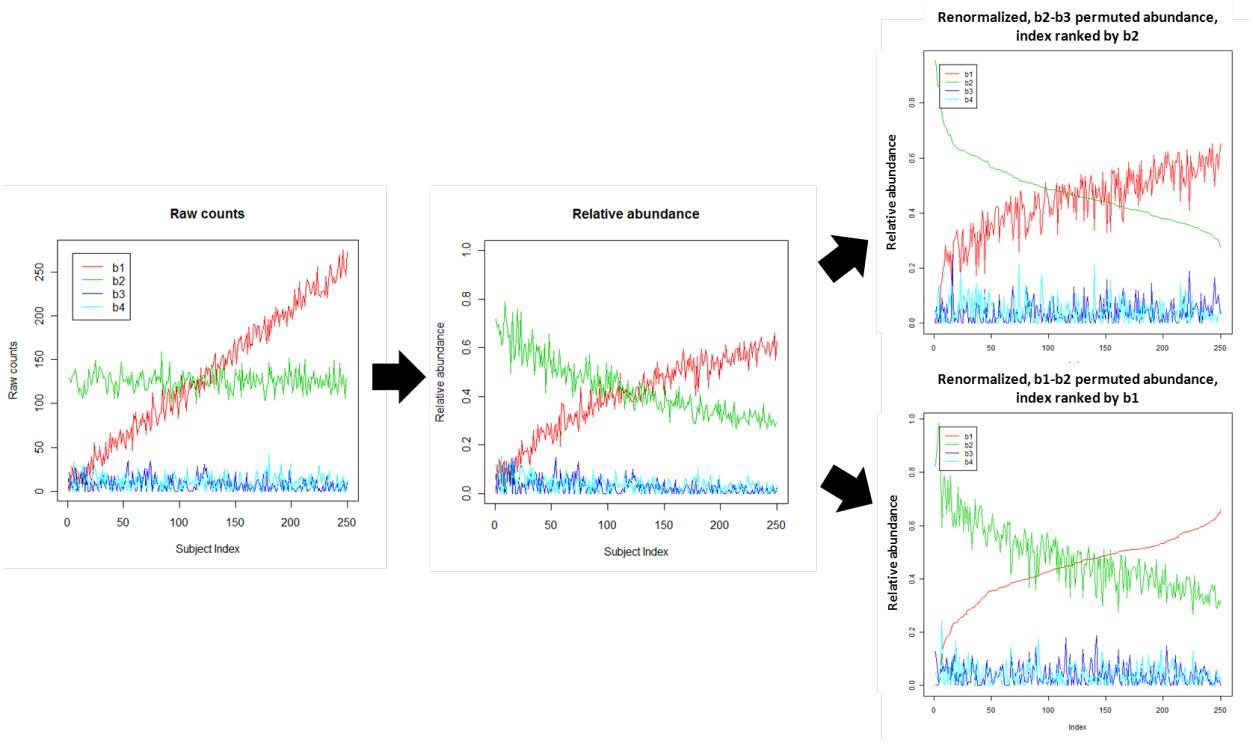


Figure SM3.5: Correlation due to compositionality alone is retained in the permutation-renormalization null distribution. Because permutation alone is insufficient to represent the null distribution of correlation due to compositionality, we recover the additional compositional structure through renormalization. Permutation of relative abundance breaks all correlation structure in the data, but renormalization of each column of permuted data reintroduces the compositional structure, by the degree to which it is present in the remaining non-permuted features' relative abundances. Correlations that exist in this renormalized permuted data are due to compositionality alone and can be used to construct an appropriate null distribution.

Bootstrap scales the confidence interval according to the overall relative abundances

As discussed above, normalization reduces the information about the absolute abundance while introducing spurious correlation. It has been shown previously [1] that the information about the absolute abundance can be retrieved through the variance/co-variance

structure of the relative abundance. Similarly, because the noise in the data (from sequencing) is assumed to be comparable between both low and high abundance microbes, normalization affects the abundance data differently based on the noise-to-signal ratio. In particular, low absolute abundance microbes are affected more strongly by the noise, thus the variance of the relative abundance is high. The reverse is true for high abundance microbes. The variance of the relative abundance directly affects the variability of the correlations between pairs of microbes, which can be observed through the variance of the bootstrap distribution of the correlations as shown through simulations below (Figure SM6). Hence, the variability of the bootstrap distribution of the correlation can be utilized as a way to recover information about the absolute abundance of the microbes.

To demonstrate the relationship between the variability of the bootstrap distribution and the absolute abundance, we performed a series of simulations in which we monitored the shape of the bootstrap distribution of correlations between two anti-correlated vectors with increasing levels of noise (Figure SM6). In particular, for each simulation, we generated a pair of perfectly anti-correlated vectors (one is <1,2,3,…,1000> and the other <1000,999,998,…,1>). Normally distributed noise with mean zero and increasing standard deviations (Figure SM6A: 50, B: 100, C: 1000, and D: 10000) were added to the two vectors, and a bootstrap correlation distribution was generated by bootstrap sampling for each of the variance (noise) levels. Although our simulation holds the absolute abundance constant and varies the noise level, this is equivalent to holding the noise level constant and varying the absolute abundance

magnitude. As shown in Figure SM6, as signal-to-noise increased, the bootstrap distribution widened, and the center shifted toward zero, which is the desired characteristic.

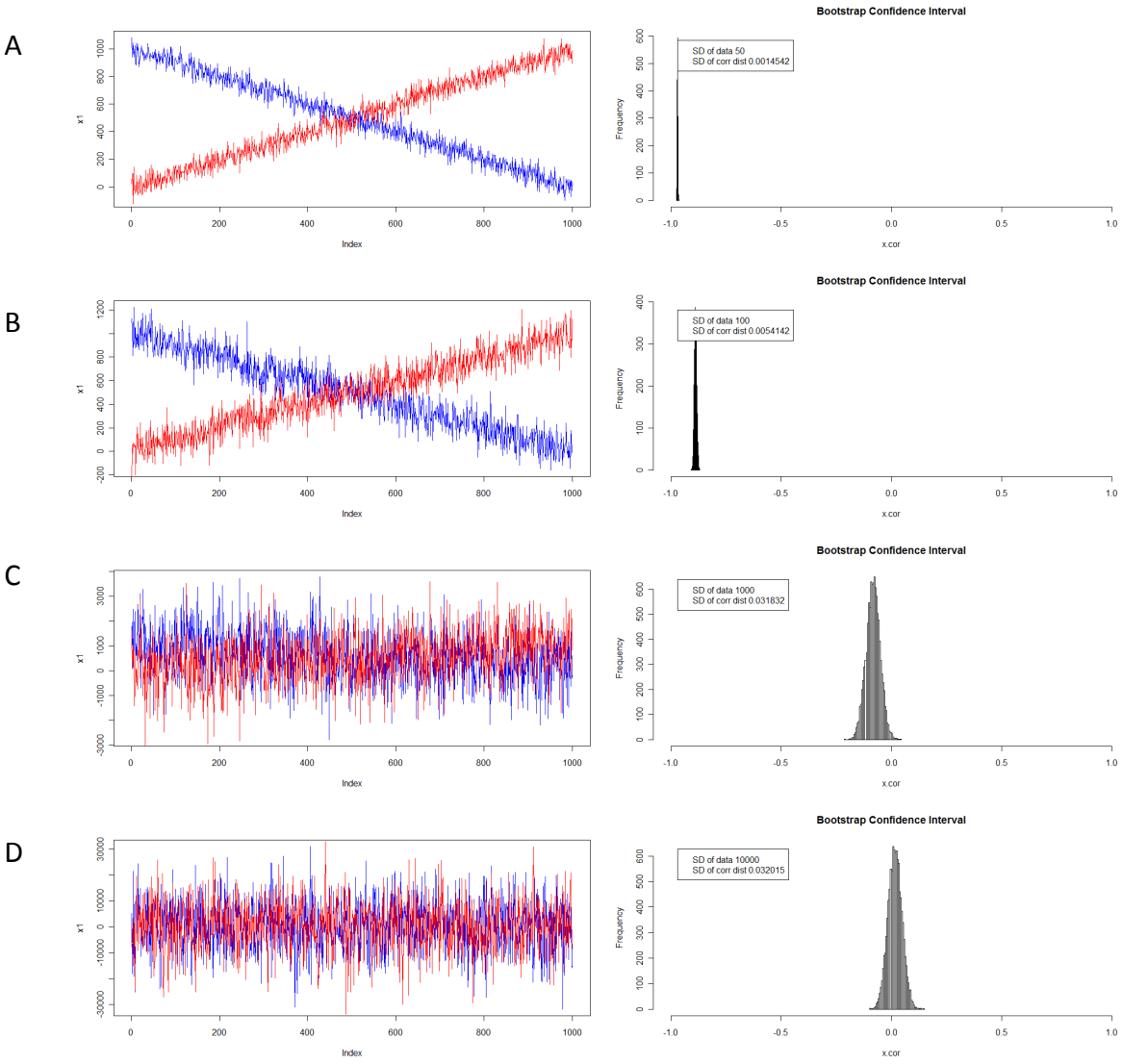
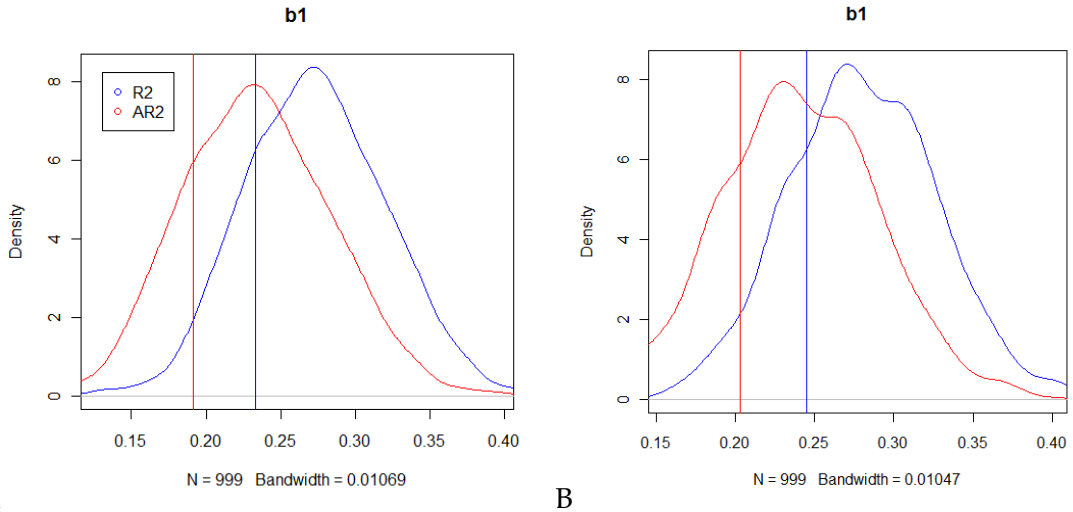


Figure SM3.6: The confidence interval provided by bootstrap assessment of correlations in compositional data scales by the signal-to-noise ratio, and is thus informative for excluding unreliable low-abundance signals most affected by compositionality. A series of pairs of vectors were generated with perfect negative correlations (one is <1,2,3,…,1000> and the other <1000,999,998,…,1>). Normally distributed noise with mean zero and increasing standard deviations (A) 50, B) 100, C) 1000, and D) 10000) were added, and a bootstrap correlation distribution was generated by bootstrap sampling for each of the variance (noise) levels. As signal-to-noise increased, the bootstrap distribution widened, and the center shifted toward zero.

Filtering based on bootstrap confidence interval of R^2 prevents over-fitting in GBLM

As an additional consideration unrelated to compositionality, we extended our application of the ReBoot procedure to GBLMs to prevent overfitting of these high-dimensional models. It has been shown that the boosting procedure for linear models can over-fit in noisy data [2], and in the case of our microbial relative abundance data, low abundance microbes can have substantially noisy data and lead to over-fitting of GBLMs. In assessing significance of GBLMs, we produced bootstrap distribution of adjusted R^2 to be compared with the null distribution, the adjustment of which provides a first step favoring simpler models with fewer parameters. However, through simulations, we observed that the bootstrapped adjusted R^2 could in many cases remain elevated, making the bootstrap confidence interval a poor estimate of the true adjusted R^2 and leading to potential false positives. **Figure SM** illustrates this in simulations, using data as in Figure SM6 and resulting in an over-fitted linear model $b1 \sim b3 +$ (12 other non-informative terms). In the over-fitted model, the bootstrap distributions were not centered at the original R^2/adjusted R^2 values, but overinflated. Note also that the bias occurs both in R^2 and adjusted R^2 estimates, as the number of predictors in the model is the same between the original model and the bootstrap models and penalizing for it as in adjusted R^2 does not correct the bias discussed here. Thus, to avoid including overfit GBLMs in later analyses, we retained only GBLMs whose bootstrap 90% confidence intervals included the true (observed) R^2 value assessed on all training data.



A B

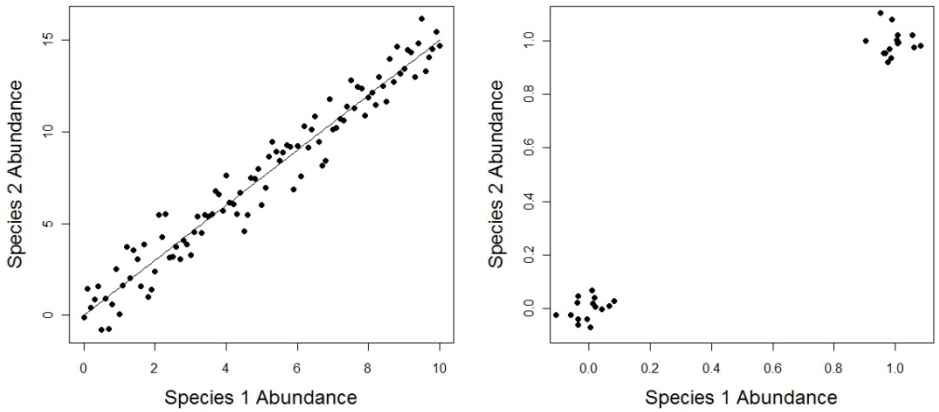
Figure SM3.7: The bootstrap confidence interval for a sparse linear model can be a biased estimator of adjusted R^2. Here, we use the data of **Error! Reference source not found.** In a GBLM predicting *b*1 relative abundance by the model: *b*1 ~ *b*3 + (12 other non-informative, low-abundance terms). R^2 (blue) and adjusted R^2 (red) based on the original data are shown as vertical lines and the bootstrap distributions are plotted as curves. (A) uses dense (nonzero) synthetic data and (B) demonstrates the case when the auxiliary predictors are sparse (>50% of the samples have zero abundance). The similarity between (A) and (B) suggests that the bias is independent of the sparsity of the predictors. In our application of GBLMs, to avoid including overfitting, we thus retained only models with bootstrap 90% confidence intervals including the R^2 observed from all training data.

A strategy to evaluate ReBoot procedure

In Faust et al. (2012), we have proposed a procedure to account for compositionality, the induction of spurious correlation by normalization. Called ReBoot, the procedure breaks any correlation structure by permutation and re-introduces the compositional correlation structure by re-normalization. Although the application of ReBoot to the 16S relative abundance data from the Human Microbiome Project yielded biologically sensible microbial co-occurrence network, a formal test of validity and analysis of the performance of ReBoot under different scenario is still needed.

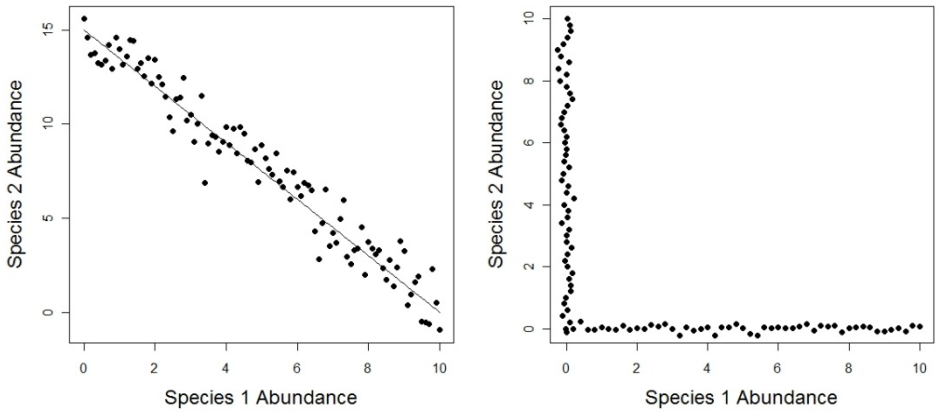
To assess the performance of ReBoot, we need both a synthetic (computationally simulated) gold standard and a biological gold standard. One of the biggest challenges of simulating relative abundances is simulating the noise. The noise present in microbial abundance data arises from two sources: technical noise and biological noise. The technical noise is a result of sample preparation and sequencing. This has been well explored in the sequencing literature and can be modeled by lognormal distribution. That is, in the absence of biological variation, microbial abundance can be simulated by a lognormal distribution.

Microbes may have natural variation across body sites and across individuals, for example Bacteroides can varies from 5-95% in the stool samples. This biological variation may correlate with variation of another microbe, hence a co-occurrence. To simulate this co-variation pattern, we may "spike-in" different pattern of variation. Four major co-variation patterns exist: 1) positive co-variation (e.g. the end product of metabolism of one can be used by the second) 2) co-occurrence (e.g. two microbes are susceptible to a toxin from a third) 3) negative co-variation (e.g. two microbes competing for the same limited resource), and 4) co-exclusion (e.g. each microbe produces a toxin against the other).



1) Positive co-variation

2) Co-occurrence



3) Negative co-variation

4) Co-exclusion

Figure SM3.8: Four patterns of co-variation: 1) positive co-variation represents cases where the end product of metabolism of one can be used by the second 2) co-occurrence represents cases where two microbes are susceptible to a toxin from a third 3) negative co-variation represents cases where two microbes competing for the same limited resource, and 4) co-exclusion represents cases where each microbe produces a toxin against the other.

Co-variation

We can simulate the absolute abundances of a pair of microbes jointly by a multivariate normal distribution and then add lognormal noise to the variation to simulate the technical noise. The strength of the co-variation can be directly controlled by changing the variance-covariance matrix of the multivariate normal distribution.

Co-occurrence

We first simulate the baseline abundance of the microbes by lognormal distribution. Then a subset of samples is chosen at random. The abundances of the pair of microbes in question are elevated jointly by a fix margin so as to create the high abundance cluster like that shown in Figure 1(2). The strength of the co-occurrence pattern can be controlled by the fraction of the samples chosen to have elevated abundances and the margin relative to the total abundance in the simulated system.

Co-exclusion

We can simulate co-exclusion pattern by first allow one microbe to vary freely by lognormal distribution with high mean, so as to allow for high range of biological variability. The second microbe is simulated under the null. Then a subset of samples is chosen at random. The abundances of the first and the second microbes in these selected samples are swapped so that in these samples the second microbes have high abundance and the first low. The strength of co-exclusion can be controlled by the mean of the lognormal of the first microbes and the fraction of the samples selected.

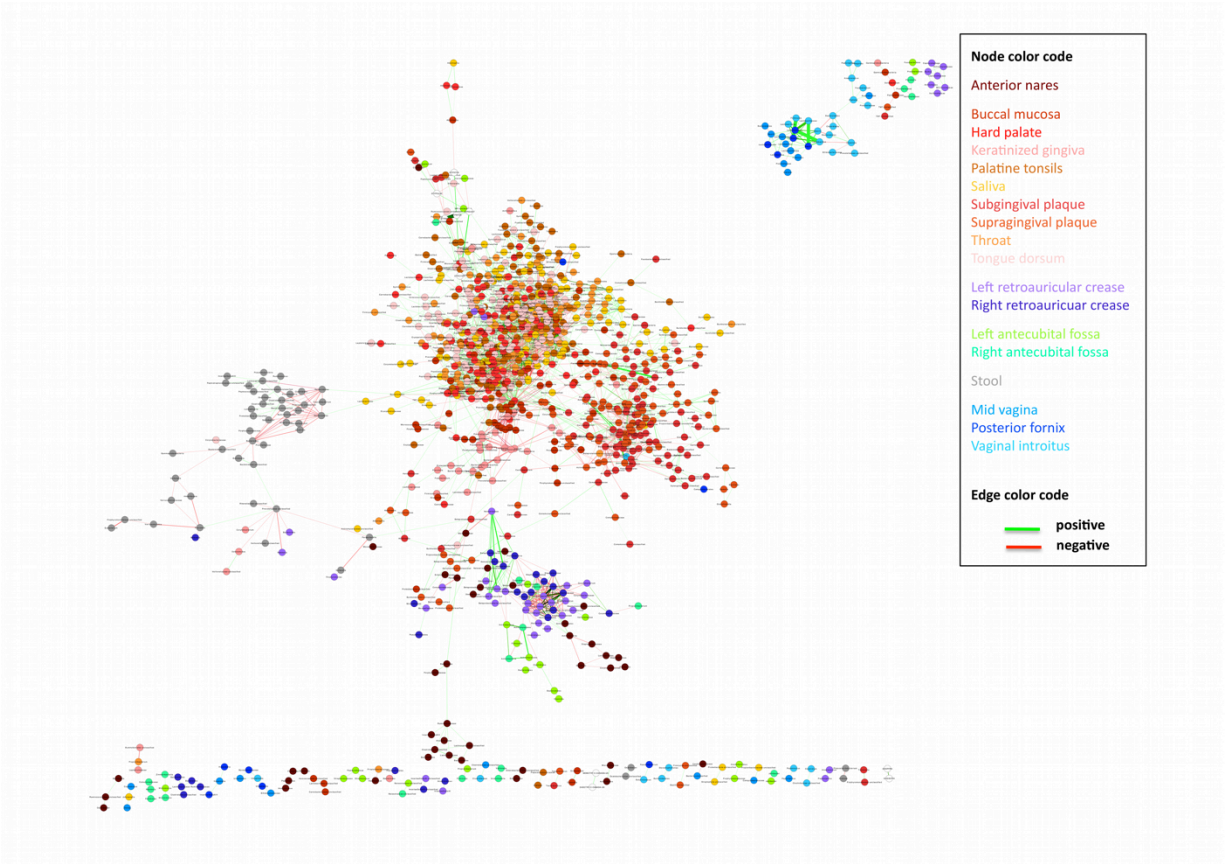
Aitchinson's log-ratio-based measure clusters with Kullback-Leibler dissimilarity

We close with a brief comment on the use of Aitchinson's proposed variance of log ratios to assess relationships between parts of compositions [1]:

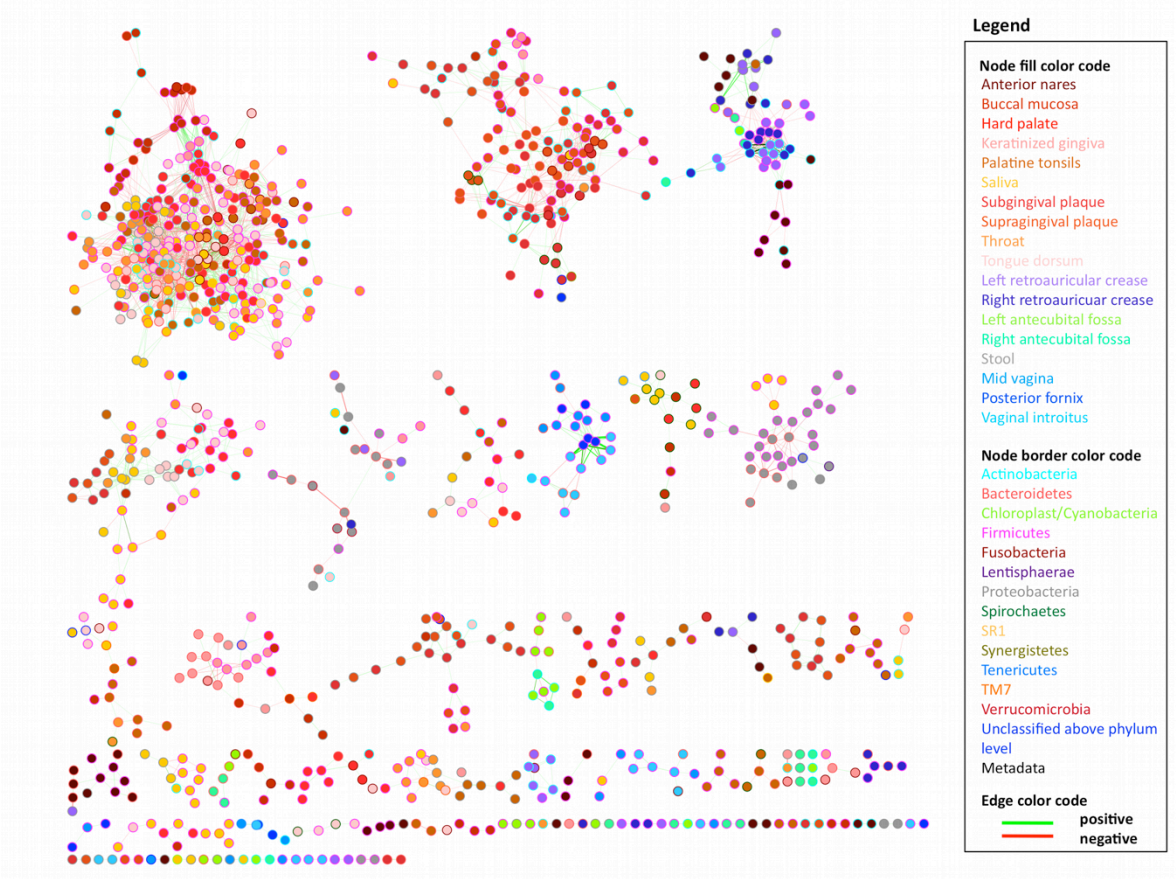
$$T = Var(\log\left(\frac{x}{y}\right))$$

This measure is zero if the ratio of two components is the same in all observed compositions and is increasing (without upper bound) with increasing variance of the ratio. The variance of log ratios can therefore be considered as a dissimilarity measure.

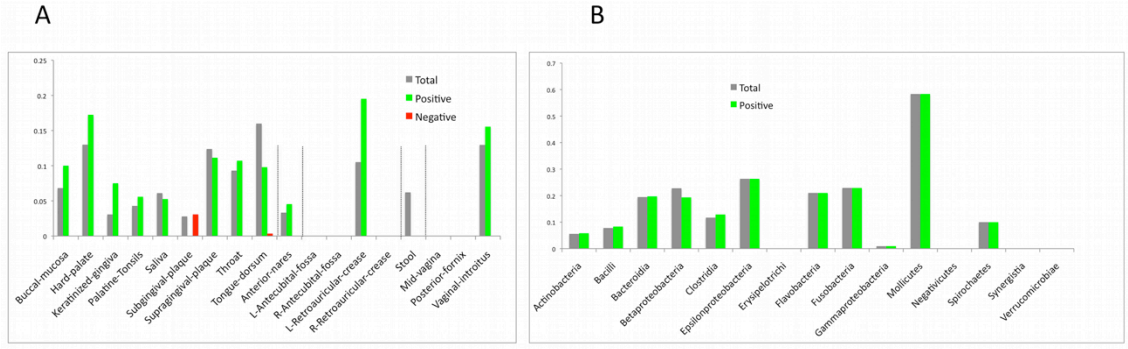
We applied the variance of log ratios to the Houston sample subset and compared the 1,000 highest- and 1,000 lowest-scoring relationships to those obtained with six other similarity and dissimilarity measures in the same data set. We found that the variance of log-ratios clustered together with other dissimilarities in terms of edge overlap, such as the Bray Curtis dissimilarity and the Euclidean distance, but shares most edges with the Kullback-Leibler dissimilarity (see Supplemental Figure 3.6). Because of its similarity to Kullback-Leibler, we did not include the variance of log-ratios in our analysis.



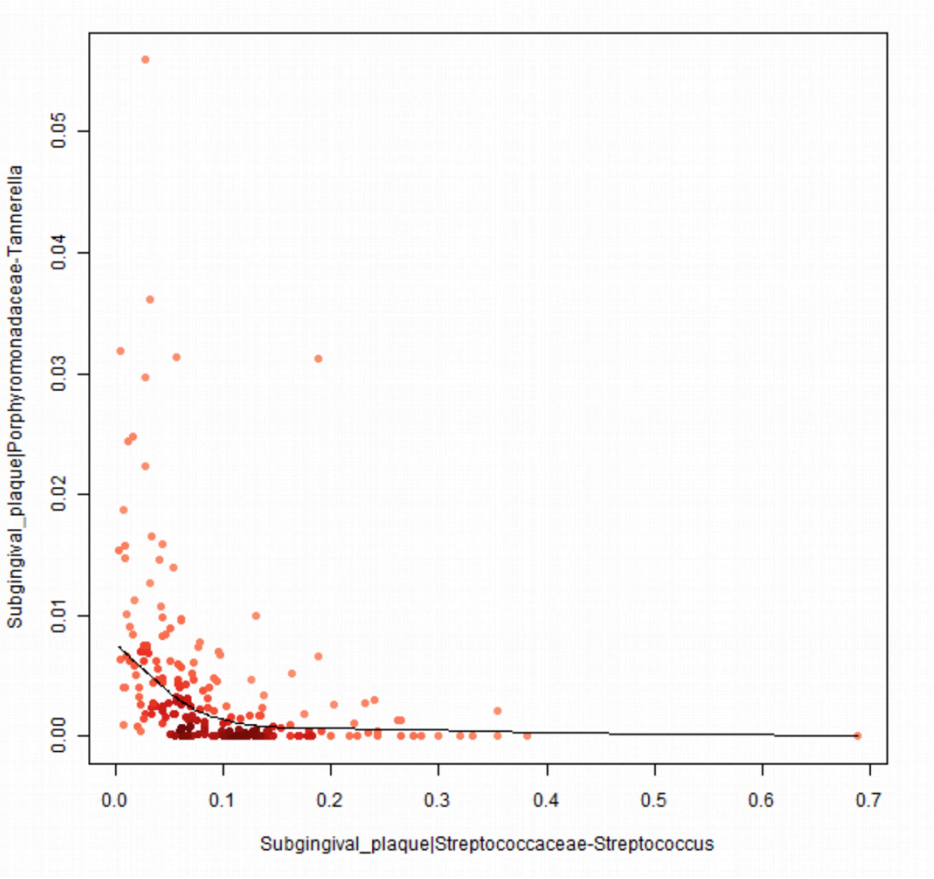
Supplemental Figure 3.1: Significant co-occurrence and co-exclusion relationships among the abundances of clades in the human microbiome. The network displays all significant phylotype associations within and across the 18 body sites sampled by the HMP. Nodes represent phylotypes (colored according to the body site in which they occur) whereas edges represent significant relationships between phylotypes. Edge thickness reflects the strength of the relationship, and edge color its directionality (green co-occurrence, red co-exclusion).



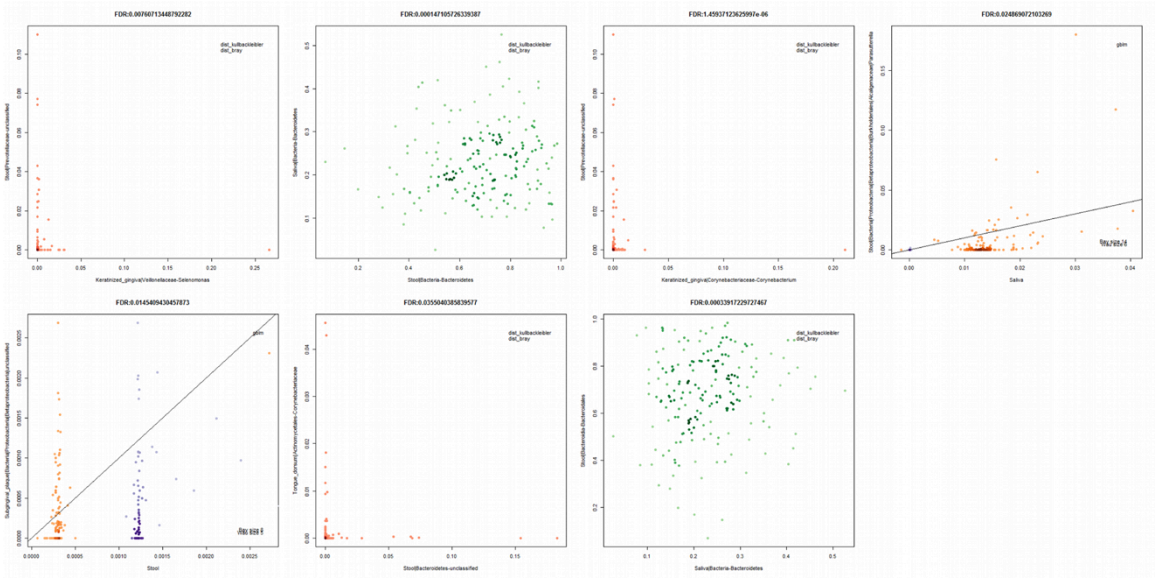
Supplemental Figure 3.2: Markov clustering of the complete phylotype network. Markov-clustered network (inflation parameter: 1.3). When clustering the cross-body site network with this inflation parameter giving optimal modularity, the network splits into the set of depicted clusters. Many of them are specific to body sites (stool, anterior nares) or areas (mouth, vagina, skin).



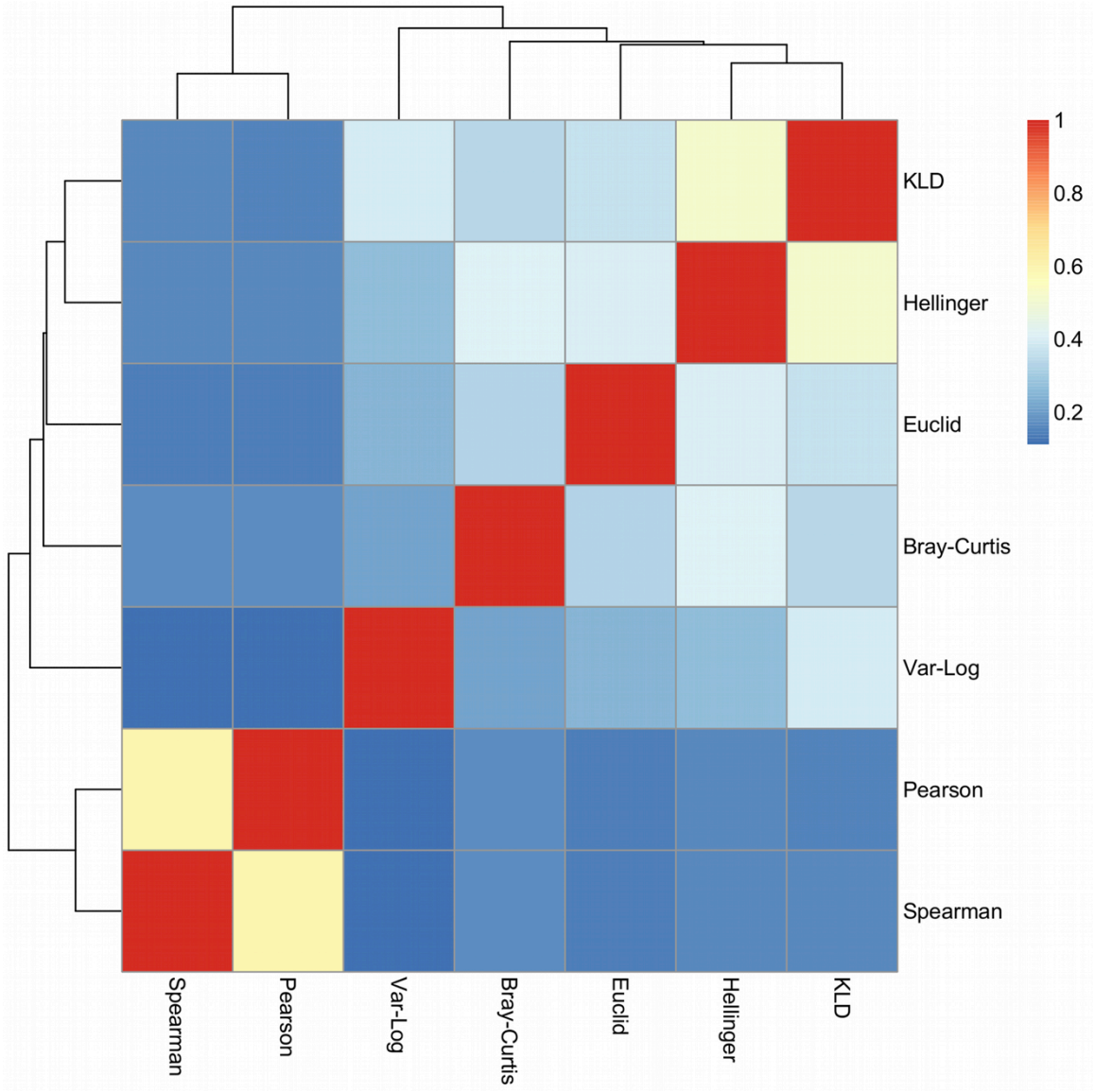
Supplemental Figure 3.3: Cluster coefficients of association networks within individual body sites and clades. Average cluster coefficients (computed with tYNA [3]) of body-site-specific (A) and class-specific (B) sub-networks. The "cliquishness" of each node within a body site or class is expressed by the average cluster coefficient, which is higher when the neighbors of each node are also connected among themselves. It can be zero if none of the nodes in the sub-network has inter-linked neighbors. The cluster coefficient was computed for all edges of a sub-network (gray bars) and for positive (green bars) and negative edges (red bars) separately. Strikingly, almost none of the negative-edge-only sub-networks had cluster coefficients above zero. In the case of the negative class sub-networks, this is a consequence of the low number of intra-class negative edges (see Figure 3E). If a negative-edge-only sub-network has a cluster coefficient of zero, it means that neighbors of a node are either not interconnected at all or that they are interconnected only by positive edges. Within the body sites, groups of phylotypes linked by negative edges likely reflect alternative communities. Members of these communities are linked among themselves by positive edges. Thus, if the positive edges are removed, the neighbors of negative nodes are no longer interlinked and the average cluster coefficient becomes zero. The high positive-edge-only cluster coefficients in classes correspond well to the high positive intra-edge number in these classes (see Figure 3E) and mean that if one member of the class is present in an individual, the other members are also likely present.



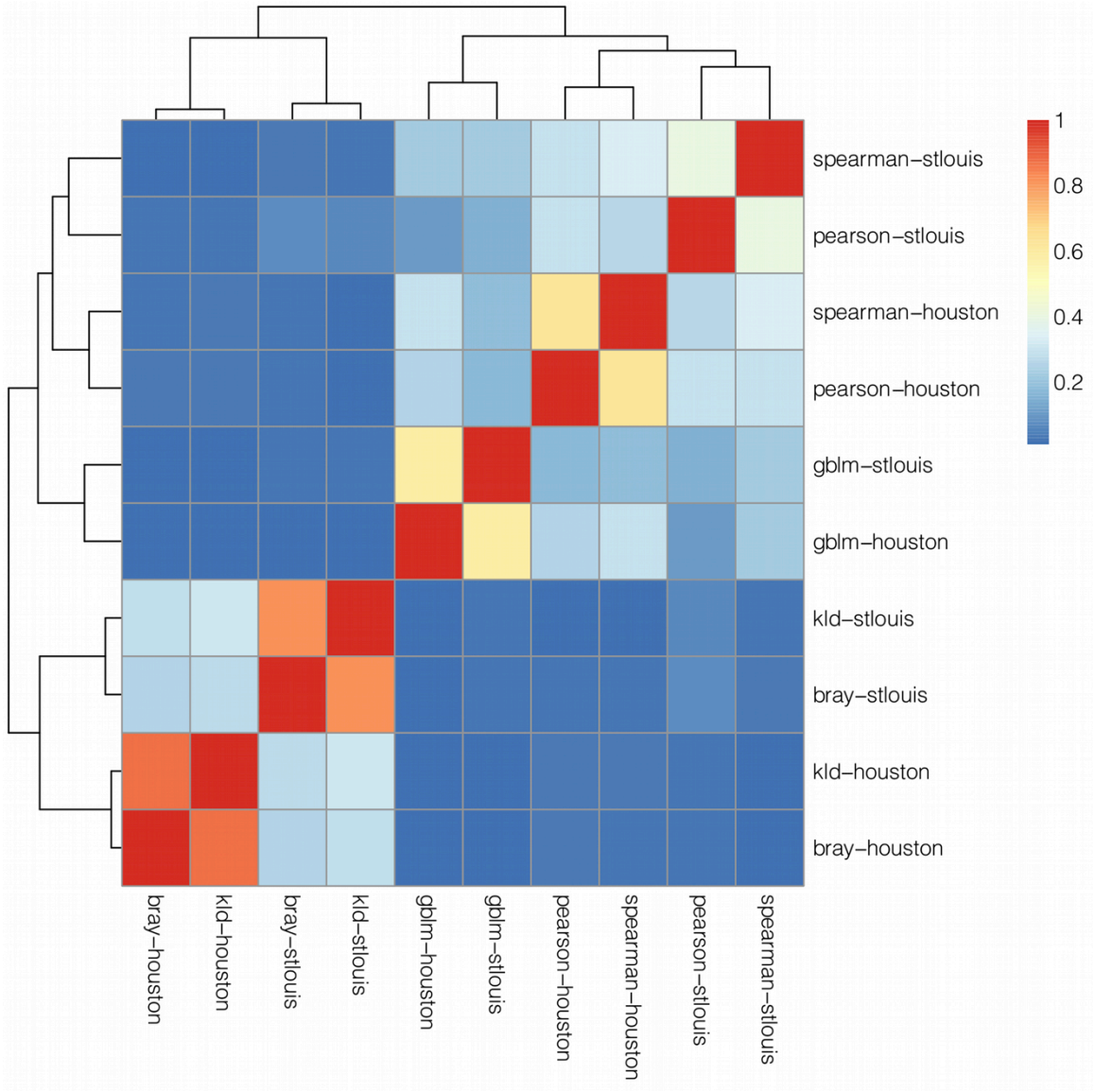
Supplemental Figure 3.4: Co-exclusion of *Tannerella* and *Streptococcus* in the subgingival plaque. The anaerobic and proteolytic *Tannerella* requires a lower pO_2 than *Streptococcus*, while *Streptococcus* is an asaccharolytic colonizer of the tooth surface that uses sugars as its primary source of carbon **[4,5]**. Between the supragingival and the subragingival plaques, as well as within the subgingival plaques, a gradient of nutrition and oxygen is present. The gradual drop of the abundance of *Tannerella* as the streptococci increase reflects the continuous nutritional and oxygen gradient between and within the supragingival and the subgingival biofilms.



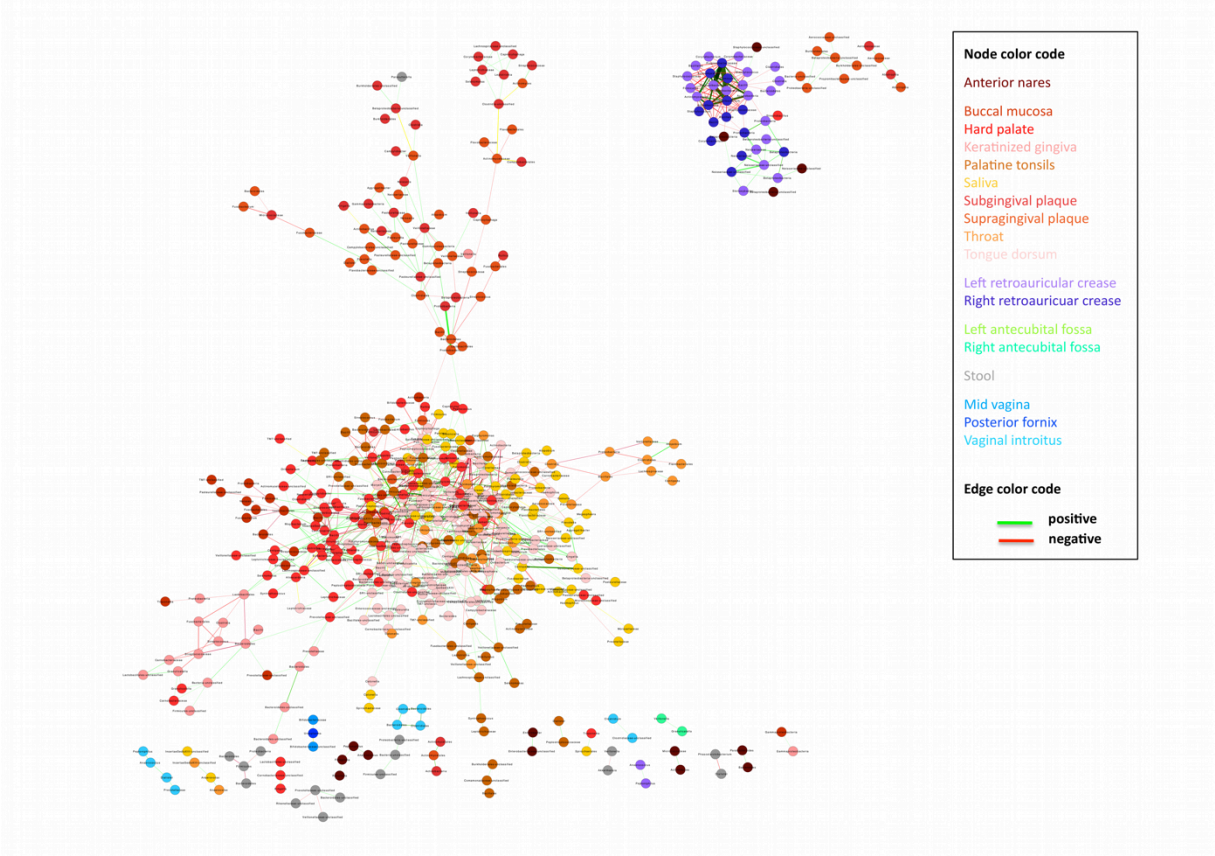
Supplemental Figure 3.5: Abundances of 18 putative associations between oral and gut microbes. Quality control plots of the raw data for all putatively significant oral/gut microbial associations showed no strong evidence for microbial transfer from the oral cavity along the digestive tract at the available level of detection. For GBLM associations, plots show predictions from the full linear model (x axis) against observed values (y axis) with the line of unity drawn as a guide, with data from the two clinical centers distinguishable by color (orange = Baylor College, purple = Washington University). None of the significant associations proved to be substantially robust from any of the nine oral body sites to gut microbes.



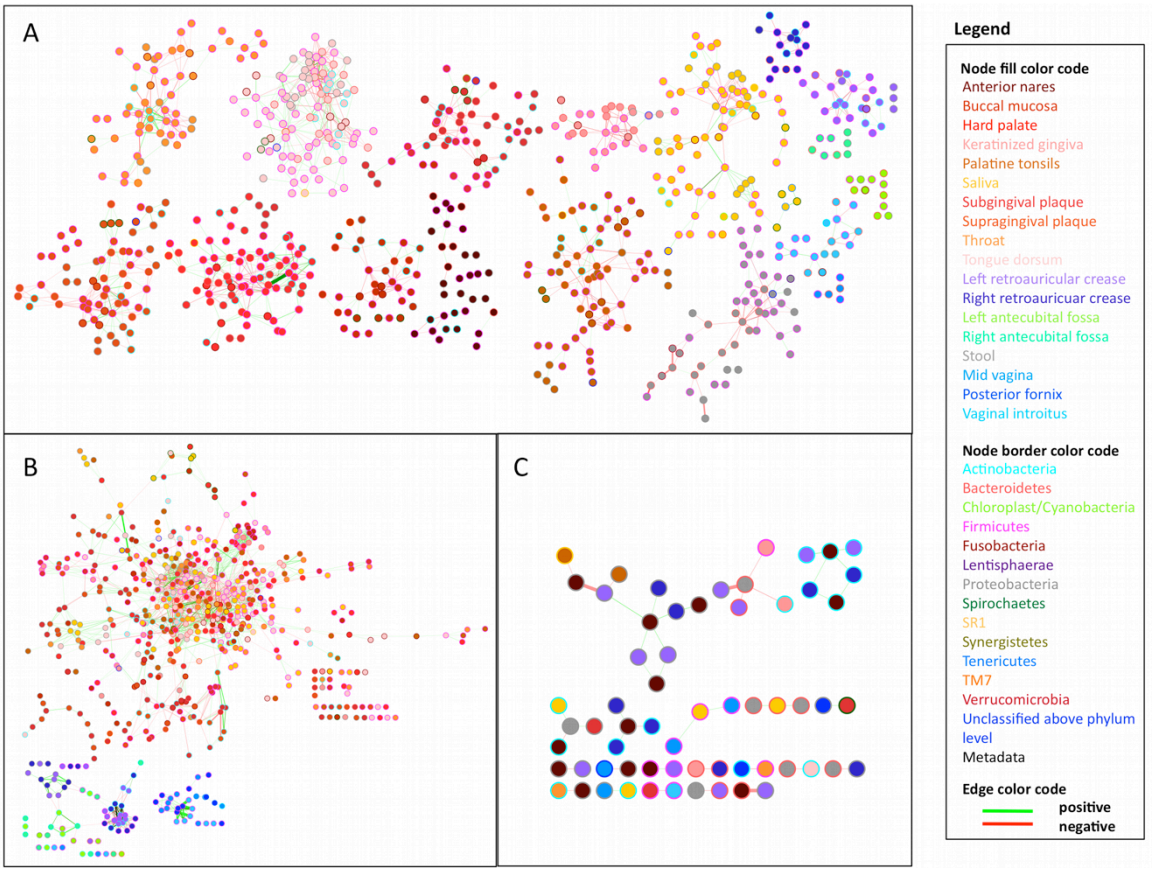
Supplemental Figure 3.6: Repeatability of network inference using seven individual similarity/dissimilarity measures with the Houston data subset. The 2,000 most extreme (1,000 top- and bottom-scoring) edges were computed for each measure in the Houston sample subset. Measure similarity was then computed as the Jaccard index of edge overlap. Abbreviations: KLD = Kullback-Leibler dissimilarity, Var-Log = variance of log-ratios, a measure recommended by Aitchison to compute associations between parts of compositions [1].



Supplemental Figure 3.7: Agreement between association networks produced by individual similarity measures and datasets. Heat map depicting the edge overlap as measured by the Jaccard index between the different methods and sample sets (Houston versus St. Louis) employed. By design from our ensemble of scoring measures, which were chosen to capture different types of microbial co-occurrences, the networks are first grouped by measure into correlations (Pearson, Spearman), GBLMs, and dissimilarities (KLD, Bray-Curtis). Each of these clusters then differentiate further according to sample set (e.g. Spearman and Pearson in Houston versus Spearman and Pearson in St. Louis).



Supplemental Figure 3.8: Intersection of networks generated independently for the Houston and St. Louis clinical center sample subsets. Our co-occurrence/exclusion network built on the combination of p-values for microbial interaction from 10 distinct networks, generated by five methods in each of two sample subsets from the HMP's Houston and St. Louis clinical centers. We examined the feasibility of treating these two clinical centers as replicates rather than semi-independent observations by performing a hard intersection, i.e. applying Simes method to each set of five methods separately and retaining only the edges significant in both. This intersection retained only 499 nodes and 938 edges, almost all of which (902, 96%) were contained in the complete network. This represents approximately 30% of the edges in the complete network, with the remainder made up of significant relationships confidently detected at only one clinical center. As the two clinical centers differed systematically in minor technical details such as input DNA concentration and chimerism during 16S sequencing [6], treating these as non-independent but non-replicate observations likely represents a more complete model of the HMP data's microbial co-occurrence and exclusion networks.



Supplemental Figure 3.9: Co-occurrence and exclusion relationships within each body site, within body areas, and between body areas. Sub-networks consisting of (A) 1,409 edges among clades within one body site, (B) 1,552 edges spanning body sites within the same area (such as the oral cavity or vagina), and (C) 44 interactions between distinct body areas.

Appendix 3 Bibliography

1. Aitchison J. A Concise Guide to Compositional Data Analysis; 2003; Girona, Spain.

2. Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning 40: 139-157.

3. Yip KY, Yu H, Kim PM, Schultz M, Gerstein M (2006) The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. Bioinformatics 22: 2968-2970.

4. Carlsson J, Iwami Y, Yamada T (1983) Hydrogen peroxide excretion by oral streptococci and effect of lactoperoxidase-thiocyanate-hydrogen peroxide. Infect Immun 40: 70-80.

5. Tanner ACR, Listgarten MA, Ebersole JL, Strezempko MN (1986) Bacteroides-Forsythus Sp-Nov, a Slow-Growing, Fusiform Bacteroides Sp from the Human Oral Cavity. International Journal of Systematic Bacteriology 36: 213-221.

6. Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One 6: e27310.

		LGRC	Bhattacharya	Steiling
COPD	**Male**	93	9	52
	Female	71	5	35
Control	**Male**	26	5	83
	Female	39	12	68
Tissue type		Whole lung	Whole lung	Bronchial brushings
Gene expression platform		Agilent Whole Human Genome 4 x 44K	Affymetrix U133 Plus 2.0	Affymetrix Human Gene 1.0 ST
Total no. of probes		19,596	54,675	19,793
No. genes (probes) mappable to SDCD genes		959 (959)	942 (1,867)	944 (948)
No. genes (probes) validated		-	264 (321)	225 (226)

Table SM4.1: Validation of SDCD genes by independent datasets

Validation of SDCD genes in independent datasets

One of the challenges of gene expression signatures is the inability to replicate in independent datasets. To assess replicabiltiy of SDCD genes, we used independent gene expression datasets from Bhattacharya [1] and Steiling [2]. Key characteristics of the datasets are summarized in Table SM4.1.

The dataset from Bhattacharya et al. includes 31 whole lung samples collected from subjects who underwent lobectomy for removal of a suspected tumor—very similar to the LGRC tissue protocol. As before, we defined COPD cases as those with FEV1 < 70% and FEV1/FVC < 0.7 and controls as those with FEV1 > 80% and FEV1/FVC > 0.7; this results in 14 COPD and 17 control samples. We then identified sexually dimorphic and differentially expressed genes (see Methods); however, due to the limited smoking exposure information in the public version of the dataset (all subjects were ex-smokers), the models were only adjusted for age. Because of a difference in the gene expression platform, we used Ensembl gene

identifiers to map Bhattacharya's expression probes to the SDCD genes. A total of 1,867 probes were mapped successfully, corresponding to 942 SDCD genes. Encouragingly, 264 (28%) of the SDCD genes included in the Bhattacharya dataset meet our criteria for being both sexually dimorphic and differentially expressed.

The dataset from Steiling et al. contains expression profiles from bronchial brushings obtained from current and former smokers with (n=87) and without (n=151) COPD. We similarly mapped probes from the dataset to SDCD genes through Ensembl gene identifiers and were able to map 948 probes, which correspond to 944 SDCD genes. In this bronchial epithelial brushing dataset, our models identified 225 (23%) of the included SDCD genes.

Description of databases of hormone regulation

Regulation by hormones, especially sex hormones, may explain sexually dimorphic expression patterns of SDCD genes. A number of databases for hormone regulation exist. The mouse and human estrogen response element database (ERE DB) [3] is a collection of high-affinity putative binding sites from a computational motif scan of the human and mouse genomes. The androgen responsive gene database (ARGDB) [4] is a manually curated list of human, mouse, and rat genes with experimental evidence for regulation by androgen. A database of regulatory motifs such as JASPAR can also be used to scan for potential binding sites within a fixed window around transcription start sites (TSS's). While JASPAR motif scan is unbiased by tissue type or cell state, it is prone to false positives and sensitive to the choice of window size and motif binding strength threshold. Here we chose a conservative window size of 1kb (750bp upstream, 250bp downstream of TSS's) and a stringent p-value cutoff of 1.0e-5.

Enrichment of nuclear receptor targets from Cistrome database

Cistrome (cistrome.org/NR_Cistrome/Targets.html) is a large collection of predicted hormone response genes based on an integration of ChIP-seq experiments and gene expression profile. In addition to estrogen receptor (ESR1/2) [5,6] and androgen receptor (AR) [7,8] targets, it contains regulatory targets for progesterone receptor (PGR) [9], another important female sex hormone, vitamin D receptor (VDR) [10], and peroxisome proliferator-activated receptor gamma (PPARG) [11]. Each of the nuclear receptors in the database has a number of regulatory target lists, each corresponding to an experimental condition. For example there are three target lists for ESR1, corresponding to treatments of MCF7 cell line with estradiol (E2) for 4 and 24 hours (Table SM4.2). All of the target lists correspond to up-regulation of target genes, except for one list which corresponds to down-regulation by ESR1. Enrichment analysis proceeded by Fisher's exact test, which assess significance of the overlap between genes targeted by hormone receptors and SDCD genes. Here, hormone receptor targets were defined by threshold on Cistrome's rank-product value (RP-value). Currently there is no standard on an appropriate cutoff for RP-values. However, because the RP-values correspond roughly to adjusted p-values, 0.25 is a possible cutoff. In fact, we found that the 0.25 RP-value threshold is conservative as it yields slightly fewer than expected total number of genes targeted by estrogen receptors (observed 300-400 vs. expected 500-1000 genes; personal communication with Dr. Shirley Liu; Table SM4.2A). This conservative definition of Cistrome targets drives the p-values toward the null (i.e. less significant), thus increasing our belief in the enrichment results.

A. Cistrome

Nuclear Receptor	SDCD genes targeted	Total genes targeted	Target enrichment p-value	Cell type	Condition
Estrogen (ESR1)	21	278	0.2197	MCF7	E2, 4hr
Estrogen (ESR1)	19	282	0.4062	MCF7	E2, 24hr
Estrogen (ESR1)	25	309	0.1226	MCF7	E2, 24hr, down regulation
Estrogen (ESR2)	39	431	0.0188	U20S	E2, 16hr
Androgen (AR)	17	241	0.3307	abl	DHT, 4hr
Androgen (AR)	16	210	0.2270	abl	DHT, 16hr
Androgen (AR)	11	255	0.9196	abl	siAR
Androgen (AR)	19	262	0.0240	LnCaP	DHT, 4-6hr
Androgen (AR)	27	188	0.0082	LnCaP	DHT, 16hr
Androgen (AR)	11	287	0.9706	LnCaP	siAR
Androgen (AR)	18	237	0.2233	LnCaP	R1881, 18hr
Androgen (AR)	30	375	0.1172	LnCaP	R1881, 48hr
Progesterone (PGR)	18	362	0.5945	MCF7	P4, 3hr
Progesterone (PGR)	40	297	0.0005	MCF7	P4, 24hr
PPAR-gamma	32	371	0.0506	Adipocyte	siPPARG
PPAR-gamma	37	416	0.0268	Adipocyte	Mature vs premature
Vitamin D (VDR)	42	546	0.1322	GM10855	Calcitriol, 36hr
Vitamin D (VDR)	42	460	0.0136	GM10861	Calcitriol, 36hr

B. Other databases of hormone responsive elements

Nuclear Receptor	SDCD genes targeted	Total genes targeted	Target enrichment p-value	Database	Experiment
Estrogen (ESR1/2)	217	2980	0.0470	ERE DB [3]	Motif scan
Estrogen (ESR1)	20	295	0.3956	JASPAR [12]	Motif scan
Estrogen (ESR2)	14	209	0.4100	JASPAR [12]	Motif scan
Androgen (AR)	103	1344	0.0485	ARGDB [4]	PubMed manual curation
Androgen (AR)	21	206	0.0178	JASPAR [12]	Motif scan

Table SM4.2: List of nuclear receptor regulatory targets and their enrichment for SDCD genes.

Sensitivity Analysis

Excluding clinical anomalies

One of the fundamental requirements of a case-control analysis is the reliability of subjects' clinical classification. Heretofore in this study we have made an effort to adjudicate based on the validity of the clinical diagnosis of the LGRC cohort, excluding samples with questionable phenotypes from the analysis. Even so, we acknowledge that the remaining samples still have attributes with the potential to affect the results of our analysis. For example, in the LGRC, a subset of the subjects have clinical phenotypes considered atypical of COPD or control patients, namely the controls with unusually low diffusion capacity (DLCO) and subjects with no reported history of smoking (see Table 1). To assess the robustness of our analyses, we performed sensitivity analyses, rerunning our models (Equations 1-4) excluding samples from subjects that may be misclassified, thus assessing the robustness of the SDCD genes and their functional enrichments.

There are 26 control samples with DLCO < 80, 24 of which had gene expression profiles and were included in the original SDCD analysis. Without the 24 control samples with low DLCO, the analysis yielded 776 genes with sexually dimorphic differential expression, of which 521 were in the original SDCD gene set (overlap Fisher's exact $p = 5.58 \times 10^{-121}$). Functional enrichment analysis of these 776 genes recapitulated similar themes in the original SDCD analysis such as cell chemotaxis (GO:0050920, p=0.003), inflammatory response (GO:0050729, p=0.048), and ATP metabolic process (GO:0006200, p=0.006), but also highlighted new concepts such as DNA damage response (GO:0006977, p=0.004) and checkpoint (GO:0031571, p=0.003).

About 15% (N=37) of subjects self-reported never having smoked cigarettes, 32 of which had gene expression profiles and were included in the original SDCD analysis. Because cigarette smoking is an important risk factor of COPD, inclusion of never-smokers may obscure the effect of cigarette smoking. Thus we repeated the SDCD expression analysis without these never-smokers and identified 2,277 SDCD genes, of which 790 are in the original SDCD gene set (overlap Fisher's exact $p < 1x10^{-120}$). Despite a much higher number of identified genes, functional enrichment results again reiterate a similar set of key biological processes, such as cell locomotion (GO:0040013, p=0.007), immune response (GO:0045824, p=0.018), and inflammatory response (GO:0002536, p=0.016). Interestingly this analysis again also highlighted DNA damage response (GO:0006977, $p=1.95x10^{-5}$) and cell cycle checkpoints (GO:0031571, $p=1.24x10^{-4}$), similar to the results when excluding the samples with DLCO<80 (see Supplementary Table S7).

Modifying parameter choices

In classifying the SDCD genes we used an FDR cutoff of 0.25 in each stratified analysis. We assessed the appropriateness of this cutoff by repeating the analysis for several various cutoff values. At each cutoff we determined the significance of the overlap between the COPD- and sex-stratified analyses by Fisher's exact test. As shown in Figure SM4.1, the significance of the overlap optimizes when the FDR cutoff is between 0.1 and 0.25 (p-value $< 1x10^{-300}$) and quickly reduces beyond cutoffs greater than 0.3. Thus, the results support the use of 0.25 as the FDR cutoff.

We next used functional enrichment results to evaluate the FDR cutoff choice. For an FDR cutoff of 0.05 (135 genes identified) we found enrichment for similar functional classes as the original SDCD gene set; for instance cell-cell signaling (GO:0007267, p=0.008), hormone secretion (GO:0046888, p=0.004), and response to wounding (GO:0009611, p=0.025); however, these genes are not enriched for several key COPD processes previously found to be significant such as immune response and inflammatory response (see Supplementary Table S7E). This suggested a more liberal cutoff is beneficial for exploring the functional impact of sexually dimorphic gene expression in COPD.

Finally, as seen in Figure 1B and Supplementary Figure S2 many of the SDCD genes have relatively small effect size as measured by $\Delta\alpha$ and $\Delta\beta$, albeit statistically significant. Since genes with a small effect size may not be clinically relevant, we selected 291 SDCD genes in the top and bottom 5% of $\Delta\beta_{\text{COPD-control}}$ ($\Delta\beta > 0.049$ or $\Delta\beta < -0.053$) and of $\Delta\alpha_{\text{male-female}}$ ($\Delta\alpha > 0.055$ and $\Delta\alpha < -0.058$) and performed functional enrichment analysis. Consistent with the analysis using all SDCD genes, the results included cell-cell signaling (GO:0007267, $p=3.58\text{x}10^{-5}$), regulation of chemotaxis (GO:0050920, $p=2.16\text{x}10^{-4}$), negative regulation of hormone secretion (GO:0046888, p=0.003), and inflammatory response (GO:0006954, p=0.004) (see Supplementary Table S7F). This consistency suggests that the full set of SDCD genes is at least as functionally relevant as the genes with the largest effect sizes. The number of SDCD genes identified in each sensitivity analysis is given in Table SM4.3, and the lists of genes and their functional enrichment can be found in Supplementary Table S7.

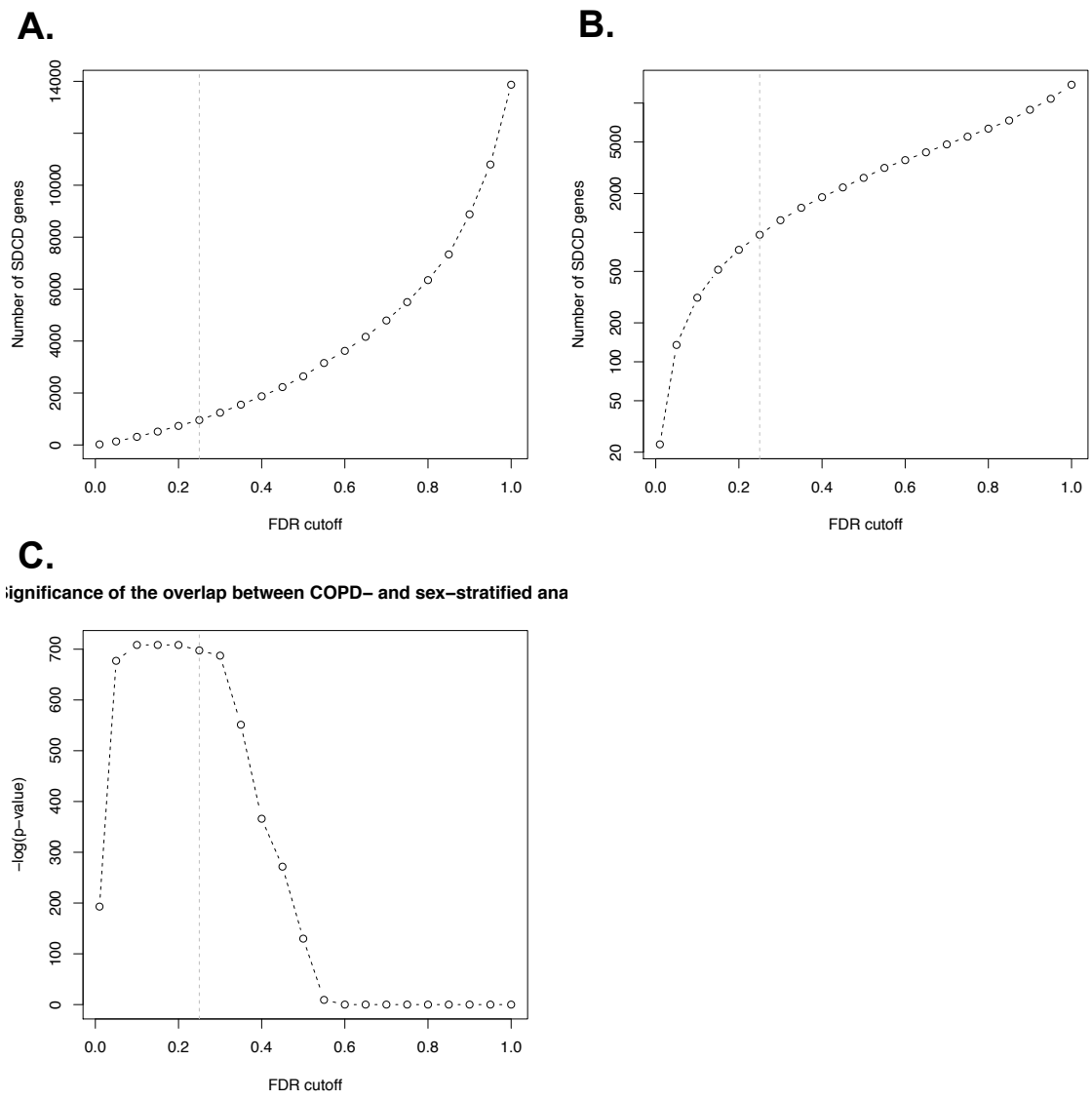


Figure SM4.1: The effect of changing FDR cutoff on the number of SDCD genes and the significance of the overlap between COPD- and sex-stratified analyses. Vertical dotted lines are at FDR = 0.25. (A) As expected, the number of SDCD genes increases as we relax the FDR cutoff. (B) The increase, when plot on the log scale, has an inflection point at around the cutoff of 0.25. (C) The significance of the overlap between the two stratified analyses peaks at cutoff between 0.05 and 0.25, which are the accepted range of the FDR cutoff.

Sensitivity Analysis	No. of subjects (M-COPD/M-Ctrl/ F-COPD/F-Ctrl)	SDCD genes identified in this analysis	No. overlap with the SDCD genes (%)
Original SDCD analysis	93/26/71/39	959	–
Without low DLCO controls	93/14/71/27	776	521 (54.33)
Without nonsmokers	90/19/68/20	2277	790 (82.37)
FDR cutoff of 0.05	–	135	135 (14.08)
Top & bottom 5% of $\Delta\beta$ and $\Delta\alpha$	–	291	291 (30.34)

Table SM4.3: Number of SDCD genes identified by various sensitivity analyses

Appendix 4A Bibliography

1. Bhattacharya S, Srisuma S, Demeo DL, Shapiro SD, Bueno R, et al. (2009) Molecular biomarkers for quantitative and discrete COPD phenotypes. Am J Respir Cell Mol Biol 40: 359-367.

2. Steiling K, van den Berge M, Hijazi K, Florido R, Campbell J, et al. (2013) A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. Am J Respir Crit Care Med 187: 933-942.

3. Bourdeau V, Deschenes J, Metivier R, Nagai Y, Nguyen D, et al. (2004) Genome-wide identification of high-affinity estrogen response elements in human and mouse. Mol Endocrinol 18: 1411-1427.

4. Jiang M, Ma Y, Chen C, Fu X, Yang S, et al. (2009) Androgen-responsive gene database: integrated knowledge on androgen-responsive genes. Mol Endocrinol 23: 1927-1933.

5. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, et al. (2006) Genome-wide analysis of estrogen receptor binding sites. Nat Genet 38: 1289-1297.

6. Vivar OI, Zhao X, Saunier EF, Griffin C, Mayba OS, et al. (2010) Estrogen receptor beta binds to and regulates three distinct classes of target genes. J Biol Chem 285: 22059-22066.

7. Yu J, Yu J, Mani RS, Cao Q, Brenner CJ, et al. (2010) An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell 17: 443-454.

8. Wang Q, Li W, Zhang Y, Yuan X, Xu K, et al. (2009) Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. Cell 138: 245-256.

9. Tang Q, Chen Y, Meyer C, Geistlinger T, Lupien M, et al. (2011) A comprehensive view of nuclear receptor cancer cistromes. Cancer Res 71: 6940-6947.

10. Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, et al. (2010) A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome Res 20: 1352-1360.

11. Lefterova MI, Zhang Y, Steger DJ, Schupp M, Schug J, et al. (2008) PPARgamma and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale. Genes Dev 22: 2941-2952.

12. Sandelin A, Wasserman WW (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol 338: 207-215.

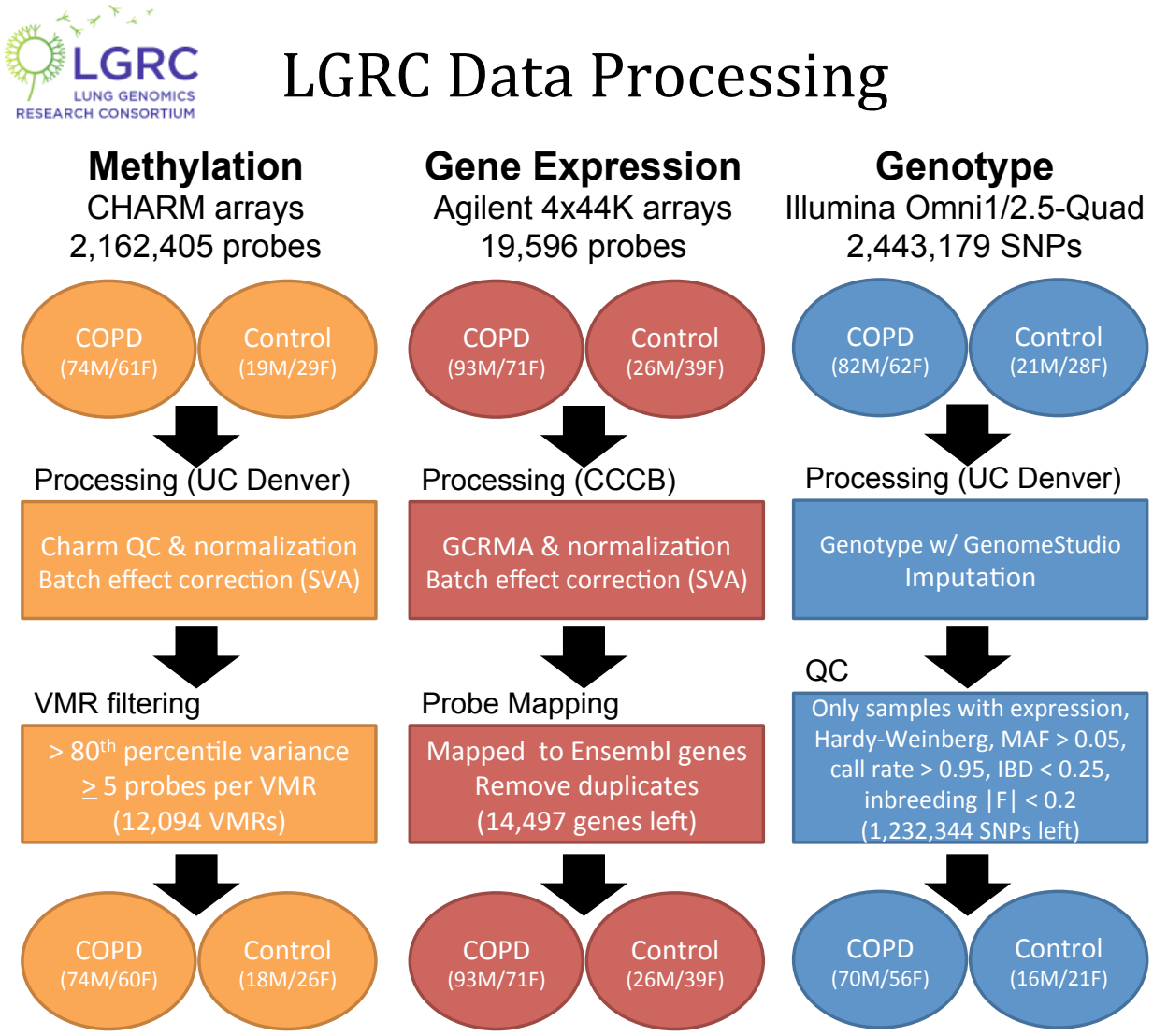


Figure S4.1: A summary of data processing for three main data types in the LGRC. Each data type was initially processed by LGRC research sites and further refined in our study. Methylation and genotype data were generated by University of Colorado at Denver while gene expression profiling was performed at University of Pittsburgh Medical Center and processed at Dana-Farber Cancer Institute's Center for Computational Cancer Biology (CCCB).

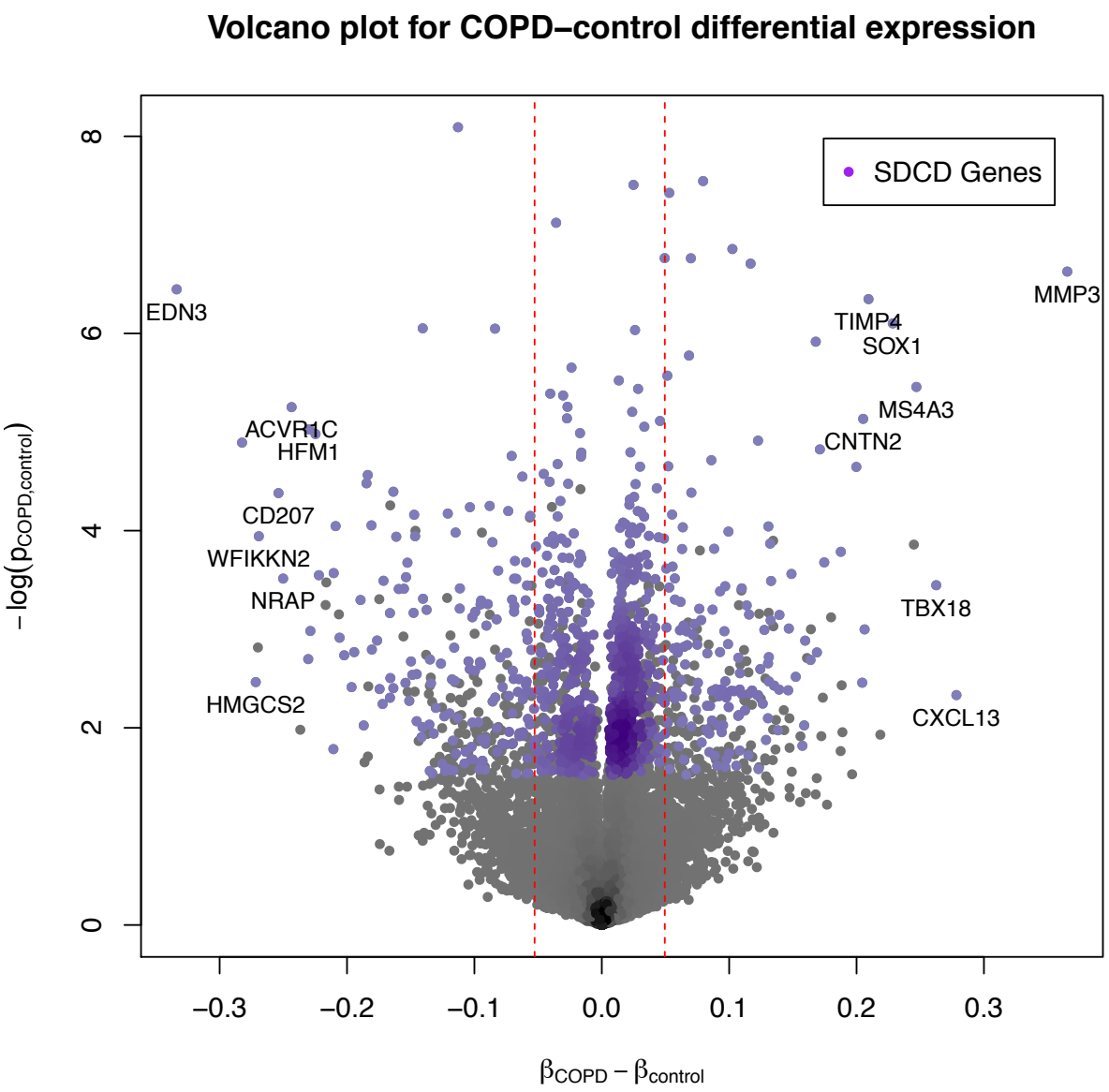
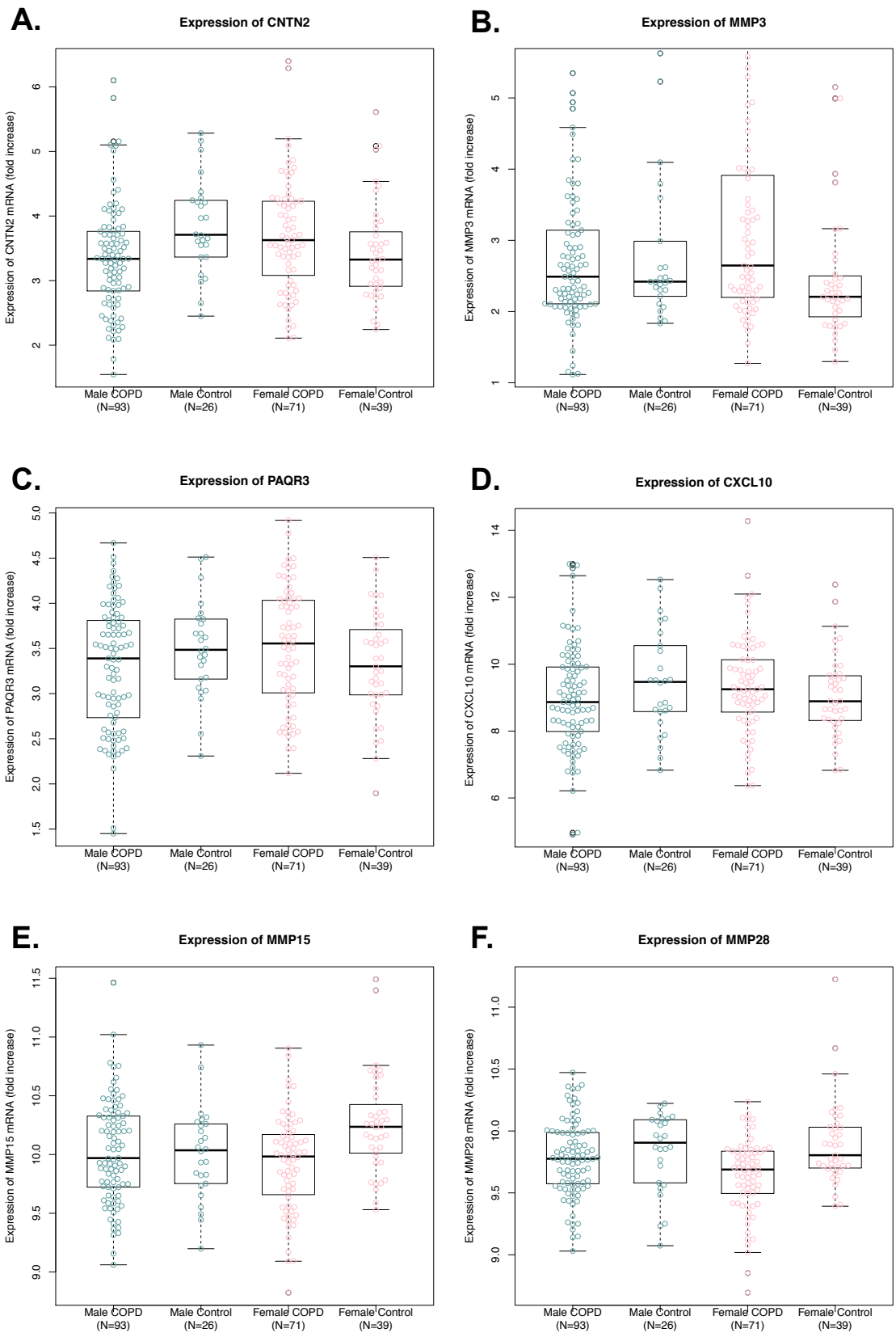


Figure S4.2: Volcano plot shows the p-values and differences of regression coefficients from the COPD-stratified SDCD analysis. The purple dots are the SDCD genes. The vertical red lines represent the 5th and 95th percentiles. The genes with large sexual dimorphic effects, as measured by the coefficient differences, include MMP3, EDN3, CXCL13, TIMP4, and CNTN2. TIMP4 and CNTN2 also have large sexual dimorphic effect as measured in the sex-stratified SDCD analysis.


A.
Expression of CNTN2
Expression of CNTN2 mRNA (fold increase)
B.
Expression of MMP3
Expression of MMP3 mRNA (fold increase)
C.
Expression of PAQR3
Expression of PAQR3 mRNA (fold increase)
D.
Expression of CXCL10
Expression of CXCL10 mRNA (fold increase)
E.
Expression of MMP15
Expression of MMP15 mRNA (fold increase)
F.
Expression of MMP28
Expression of MMP28 mRNA (fold increase)
Male COPD (N=93)
Male Control (N=26)
Female COPD (N=71)
Female Control (N=39)

(continued)

Figure S4.3: mRNA expression profile of selected SDCD genes stratified by sex and COPD status. (A) CNTN2, along with TIMP4, appear in the top corner of the volcano plot in Figure 4.1B, representing low p-values and large effect sizes. (B) MMP3 shows the largest effect size in the COPD-stratified analysis. (C) PAQR3 and (D) CXCL10 are among the SDCD genes that are sexually dimorphic in COPD but not in control samples. (E-F) In addition to MMP3, two members of matrix metalloproteinase family, MMP15 and MMP28, also exhibit sexually dimorphic and differential expression.

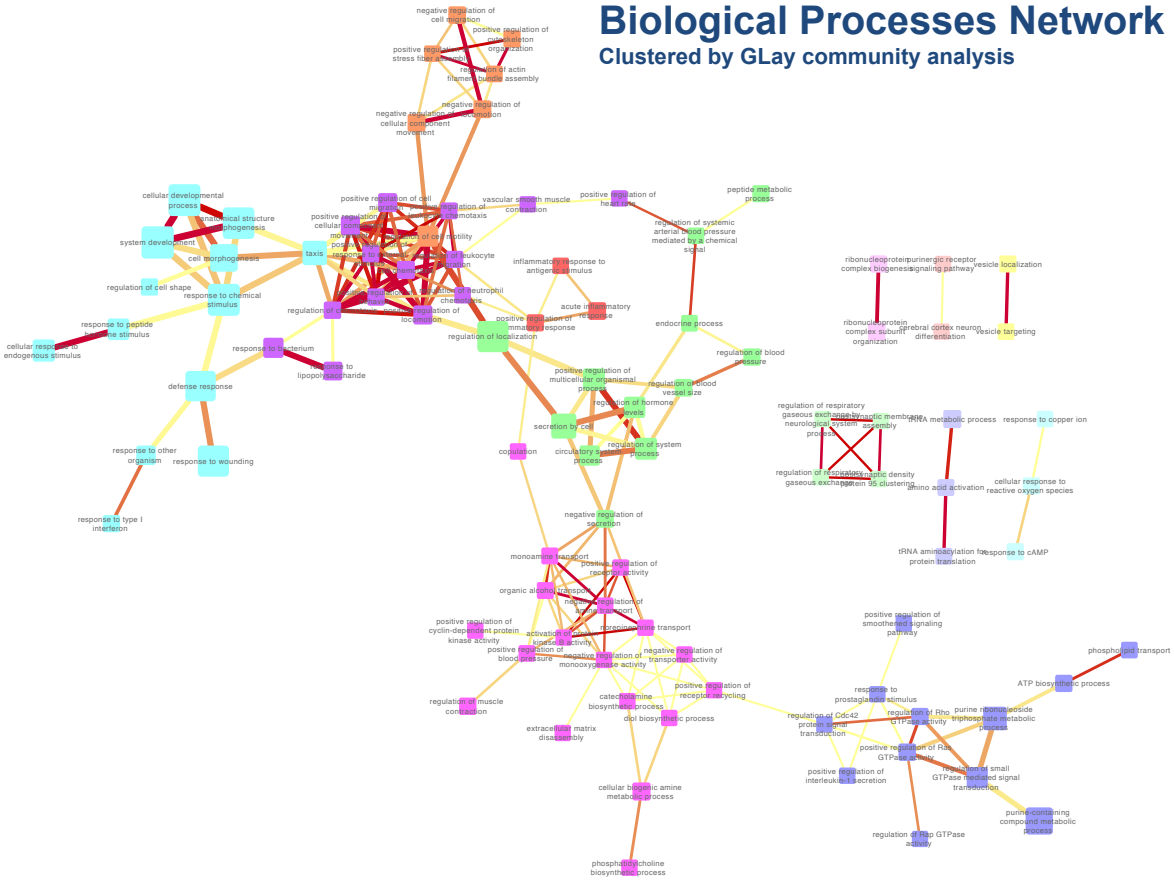
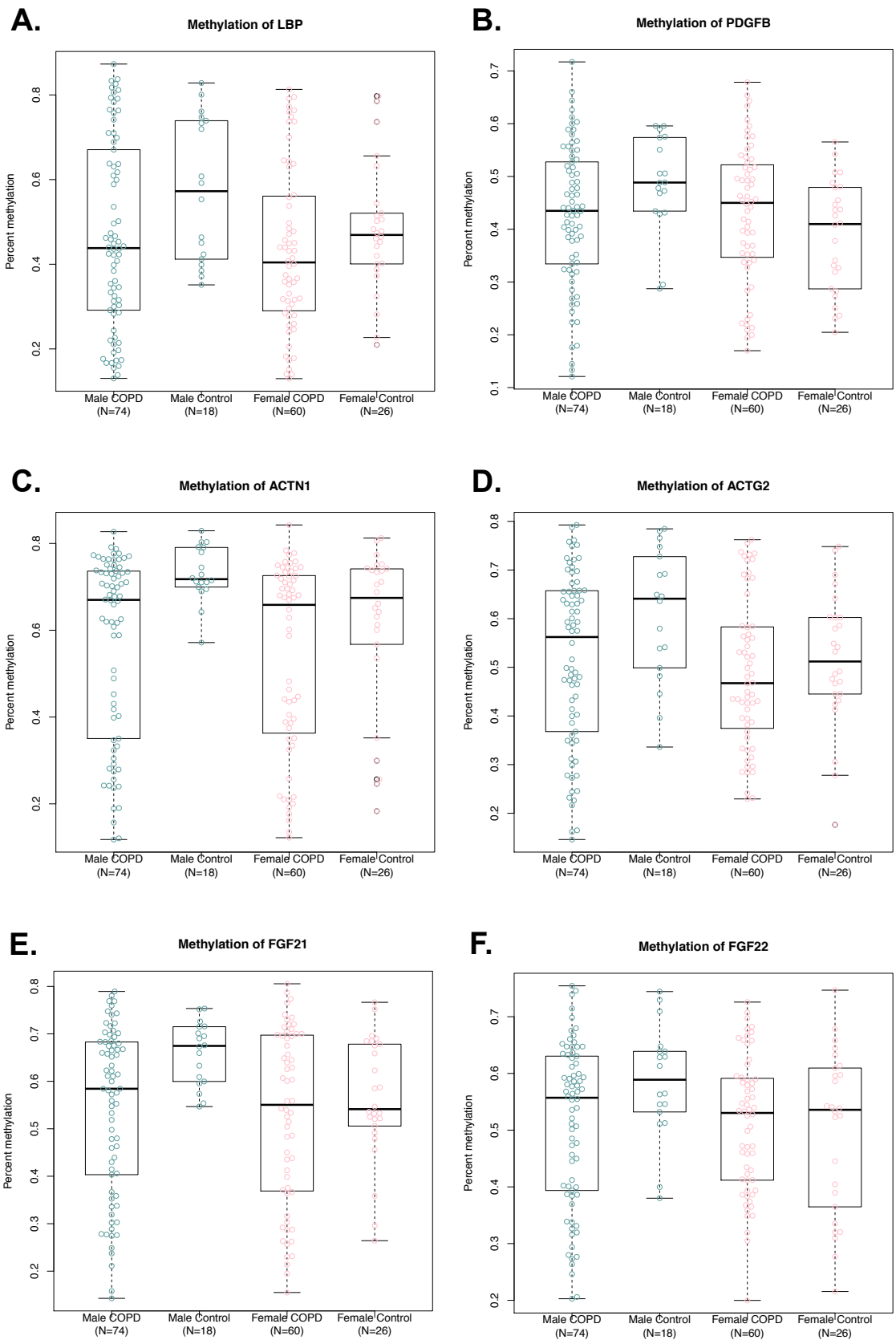


Figure S4.4: Network clustering analysis of SDCD gene-enriched biological processes. GLay community analysis of GO terms (Figure 4.2A) reveals distinct and coherent clusters of biological processes as highlighted in Figure 4.2A.

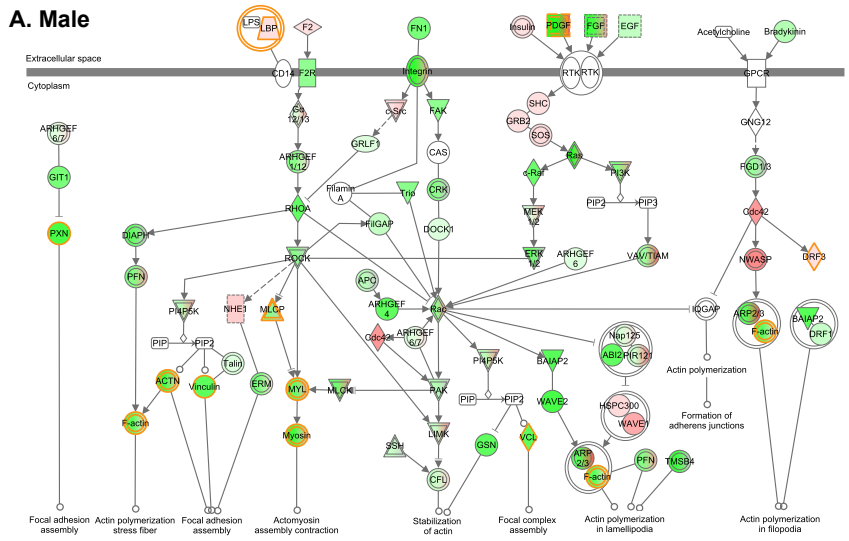

A.
Methylation of LBP
Percent methylation
0.2
0.4
0.6
0.8
Male COPD (N=74)
Male Control (N=18)
Female COPD (N=60)
Female Control (N=26)
B.
Methylation of PDGFB
Percent methylation
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Male COPD (N=74)
Male Control (N=18)
Female COPD (N=60)
Female Control (N=26)
C.
Methylation of ACTN1
Percent methylation
0.2
0.4
0.6
0.8
Male COPD (N=74)
Male Control (N=18)
Female COPD (N=60)
Female Control (N=26)
D.
Methylation of ACTG2
Percent methylation
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Male COPD (N=74)
Male Control (N=18)
Female COPD (N=60)
Female Control (N=26)
E.
Methylation of FGF21
Percent methylation
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Male COPD (N=74)
Male Control (N=18)
Female COPD (N=60)
Female Control (N=26)
F.
Methylation of FGF22
Percent methylation
0.2
0.3
0.4
0.5
0.6
0.7
Male COPD (N=74)
Male Control (N=18)
Female COPD (N=60)
Female Control (N=26)

(continued)

Figure S4.5: Boxplots showing percent methylation of VMRs proximal to SDCD genes. Stratified by sex and COPD status, these exemplify general patterns of sexual dimorphic methylation. Male control samples appear to have consistently higher level of methylation while the level of methylation in COPD and control female are largely similar. These genes included here are involved in actin cytoskeletal signaling pathway.

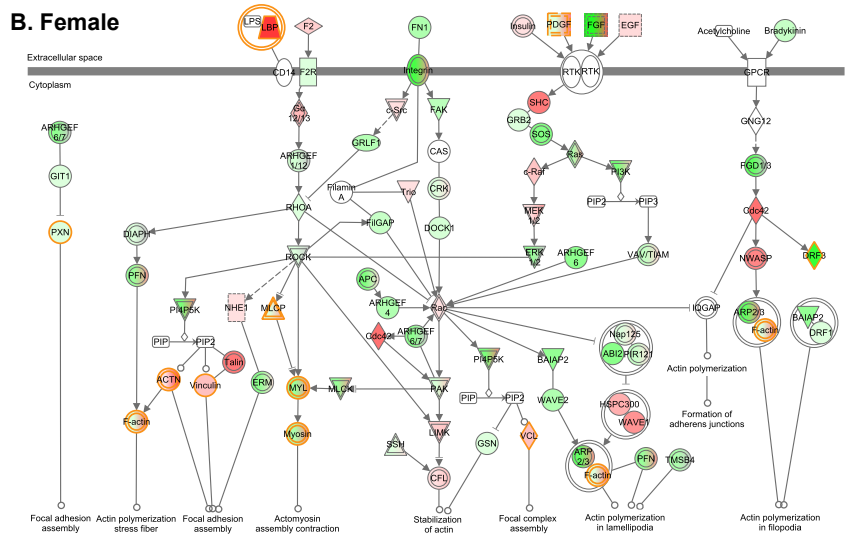
Actin Cytoskeleton Signaling

A. Male

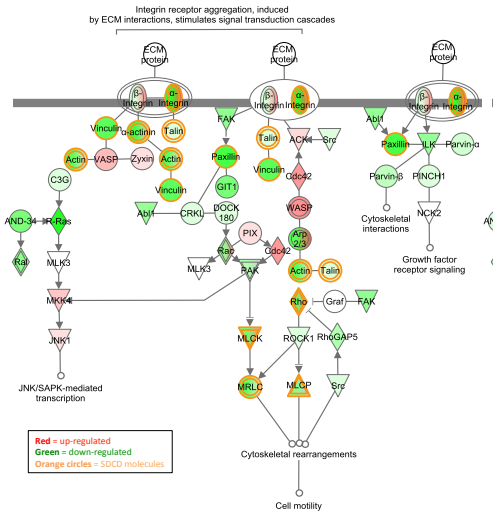


Actin Cytoskeleton Signaling

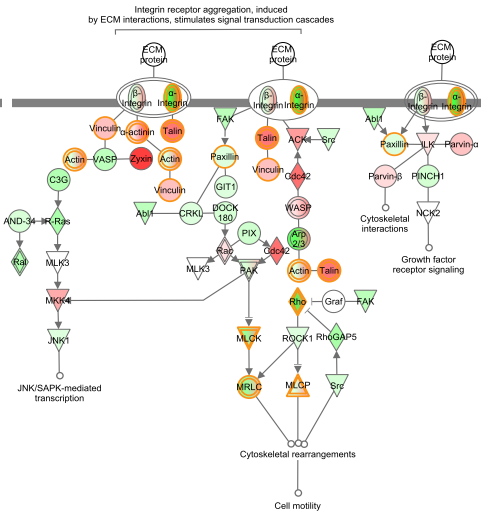
B. Female



C. Male



D. Female



(continued)

Figure S4.6: Actin cytoskeleton signaling and integrin signaling pathways represented by IPA. Node colors represent the regression coefficients: (A,C) male (α_{male}) and (B,D) female (α_{female}) (green=negative, red=positive). The genes marked by orange edges are SDCD genes. Full legend can be found in Figure 4.4. (A-B) While most molecules are similarly expressed between male and female, key growth factors and signaling molecules appear to be SDCD genes. Similar to VEGF-signaling pathway, actin, alpha acting, paxillin, and myosin activate downstream key events such as focal adhesion. (C-D) In integrin signaling pathway, the actin-alpha actin-vinculin-paxillin complex (top left) corresponds to the signaling complex of VEGF-activated cell migration pathway. VEGF and integrin signaling pathways may form a signal transduction cascade potentially impacting sex-specific features of cytoskeletal rearrangements and cell motility.

Supplemental Tables

These files are available at http://www.filedropper.com/chapter4suptables and as supplemental files to this dissertation.

Table S4.1: Number of LGRC samples with various genomic assays. [DOCX]

Table S4.2: The SDCD Genes. (A) A list of 959 SDCD genes. (B) Subset of 150 SDCD genes that are sexually dimorphic in COPD but not in controls and (C) their functional enrichment from DAVID (david.abcc.ncifcrf.gov). [XLSX]

Table S4.3: Top SDCD genes that are related to COPD. [DOCX]

Table S4.4: Replication of SDCD genes by independent datasets. SDCD genes replicated in (A) Bhattacharya et al. dataset and (B) Steiling et al. dataset. [XLSX]

Table S4.5: Functional enrichment analyses of SDCD genes. (A) Functional enrichment of SDCD genes based on Gene Ontology. (B-C) Results from gene set enrichment analysis (GSEA) using a list of all genes ranked by (B) $\Delta\beta_{\text{COPD-control}}$ and (C) adjusted $p_{\text{COPD-control}}$. Only results with $p < 0.1$ are listed. (D) Functional enrichment analysis using DAVID (david.abcc.ncifcrf.gov). (E) Canonical pathway enrichment results from Ingenuity Pathway Analysis (IPA, Ingenuity Systems, www.ingenuity.com). [XLSX]

Table S4.6: Regulatory mechanisms of SDCD genes. (A) A list of variably methylated regions (VMRs) with evidence for sexually dimorphic differential methylation. For each SDCD VMR, a list of near by genes within 10kb is given, along with the key statistics from the analysis. (B) Functional enrichment analysis of the genes associated with the sexually dimorphic VMRs. (C) A list of sexually dimorphic eQTL with the key statistics from the male-only and female-only eQTL analyses. [XLSX]

Table S4.7: Sexually dimorphic differential expressed genes identified in sensitivity analysis and their functional enrichment. (A) Genes and (B) their functional enrichment from the analysis without controls with low diffusion capacity (DLCO < 80) (C) Genes and (D) their functional enrichment from the analysis without samples who self-reported never having smoked cigarettes. (E) Functional enrichment of SDCD genes identified by FDR cutoff of 0.05. (F) Functional enrichment of SDCD genes with effect sizes in the top and bottom 5% of $\Delta\alpha_{\text{male-female}}$ and $\Delta\beta_{\text{COPD-control}}$ (i.e. outside of the 90% confidence interval). [XLSX]

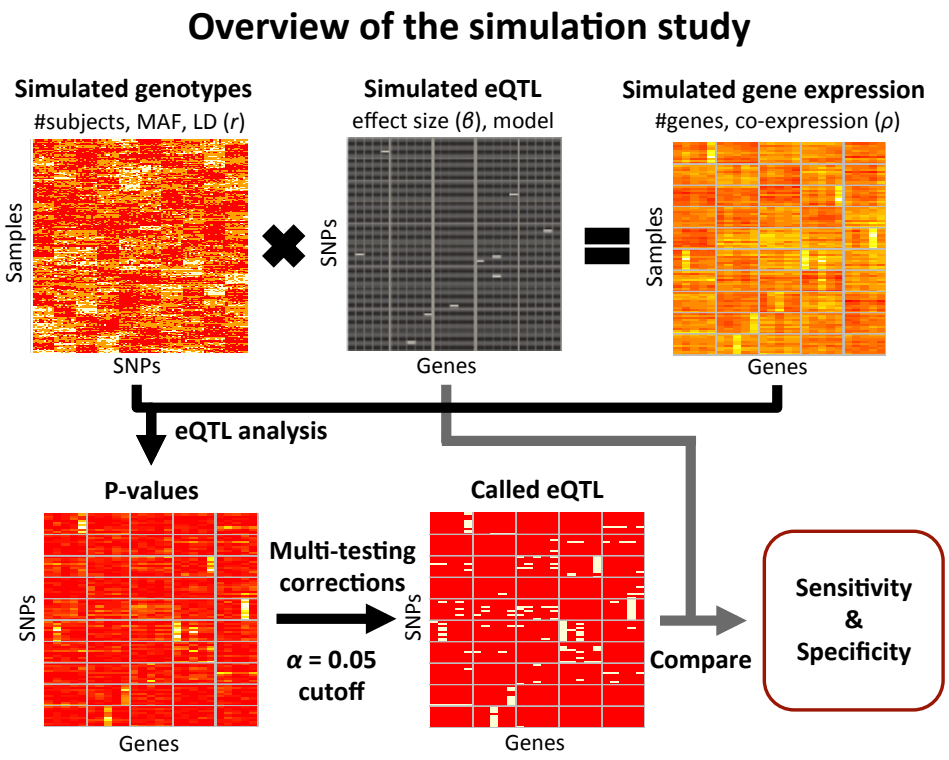


Figure S5.1: Overview of the simulation study. For each iteration of the simulation, we draw from a correlated binomial distribution to generate genotypes. Pairs of SNP and genes were selected to be "true" eQTL. Then based on the eQTL and genotypes, mean and standard deviation of expression for each gene is calculated. We use a multivariate normal distribution to simulate gene expression data. Then, eQTL analysis and multiple testing adjustment are performed, and eQTL are called using the significance level of 0.05. Comparing the called eQTL and the simulated eQTL yields specificity and sensitivity.

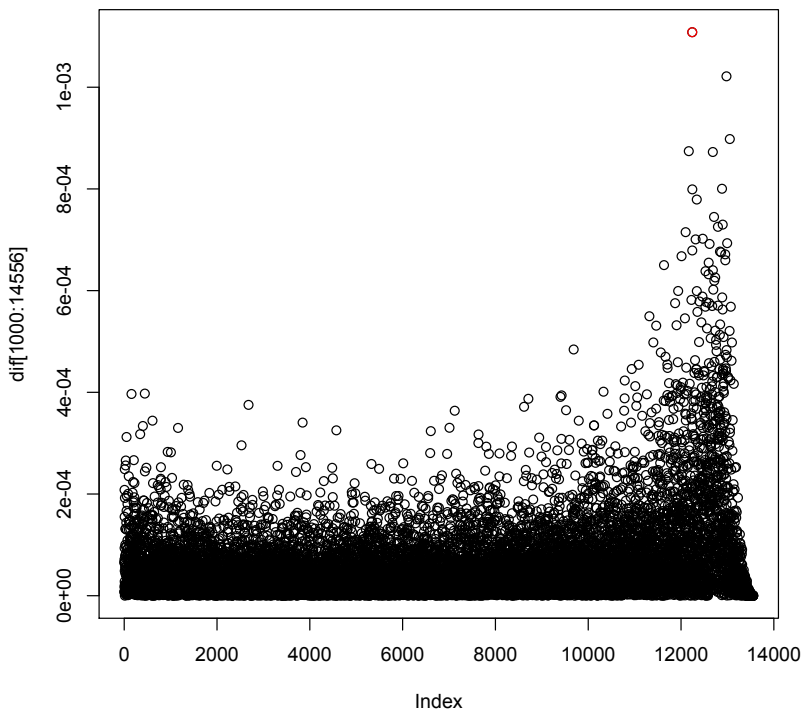
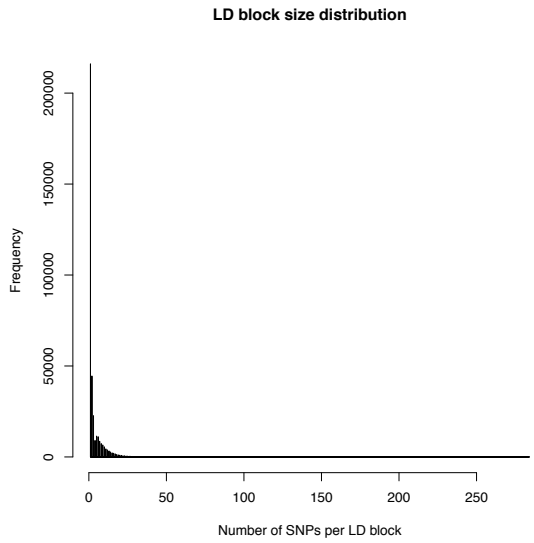
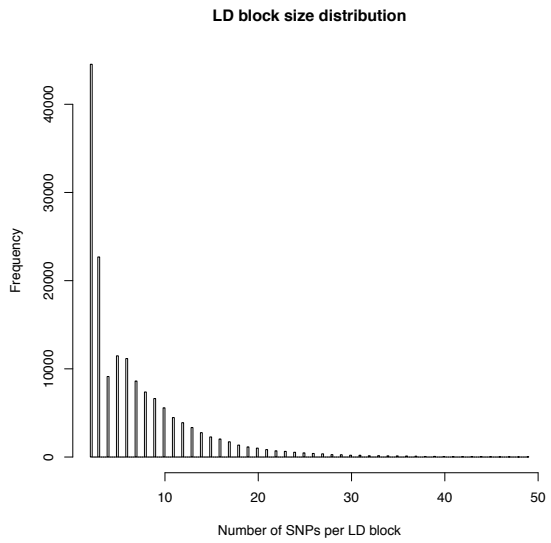


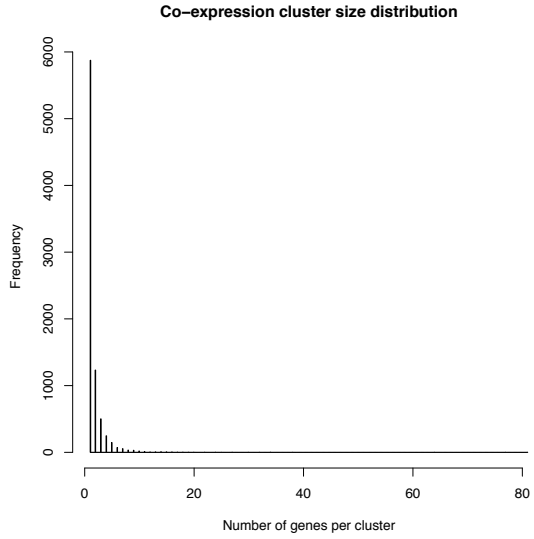
Figure S5.2: Height distances between subsequent cluster joining events from hierarchical clustering of LGRC gene expression data. Hierarchical clustering was performed with 1-|cor(X,Y)| as distance. The dendrogram was cut using the peak distance (marked in red) so as to maximize the distance between clusters. This corresponds to dendrogram tree height of 0.4, which is used to define clusters.



A.



B.



C.

Figure S5.3: Cluster size distribution. (A) LD blocks range in size from one to several hundred SNPs. (B) A truncated histogram of LD block sizes (1 < NSNPS < 50). (C) Size distribution of the co-expression clusters identified by hierarchical clustering.

SNP			Gene			eQTL		
SNP	Chr	Position	Gene	Chr	TSS	Unadjusted P-value	BBER-adjusted FDR	Cis
rs2532316	17	44213712	MAPK8IP1	11	45907202	1.67E-47	3.58E-39	NO
rs2469933	17	44285531	MAPK8IP1	11	45907202	8.76E-50	3.54E-41	NO
rs2732651	17	44345063	MAPK8IP1	11	45907202	2.80E-49	1.13E-40	NO
rs17650901	17	44039691	MAPK8IP1	11	45907202	3.16E-42	3.33E-38	NO
kgp8197546	17	44041562	MAPK8IP1	11	45907202	7.77E-42	2.00E-36	NO
rs2532274	17	44247164	MAPK8IP1	11	45907202	3.26E-46	3.20E-43	NO
rs2395943	6	42940673	PEX6	6	42946958	4.49E-44	3.81E-36	YES
rs11982736	7	55855180	PSPH	7	56119297	5.44E-49	3.03E-41	YES
kgp6231899	16	24845143	SLC5A11	16	24857162	2.30E-45	1.06E-36	YES
kgp11777277	16	24856775	SLC5A11	16	24857162	4.72E-45	2.18E-36	YES
rs2550309	16	56434742	AMFR	16	56459450	4.88E-45	3.15E-36	YES
kgp11544682	16	56436824	AMFR	16	56459450	1.83E-45	1.18E-36	YES
kgp8174079	16	56442734	AMFR	16	56459450	4.88E-45	3.15E-36	YES
rs1382359	16	56399514	AMFR	16	56459450	4.40E-46	2.37E-37	YES
kgp1284210	16	56442066	AMFR	16	56459450	1.83E-45	1.18E-36	YES
kgp6813453	16	56395610	AMFR	16	56459450	4.40E-46	4.74E-37	YES
kgp2821240	16	56462000	AMFR	16	56459450	4.88E-45	4.77E-37	YES
rs4924	16	56396486	AMFR	16	56459450	2.30E-46	2.47E-37	YES
rs1478478	16	56402531	AMFR	16	56459450	4.88E-45	2.62E-36	YES
rs4251689	8	145741130	LRRC14	8	145743376	1.49E-45	1.20E-36	YES
kgp2553412	8	145764360	LRRC14	8	145743376	1.79E-43	6.58E-37	YES
rs3816732	8	145724275	LRRC14	8	145743376	6.59E-46	5.32E-37	YES
rs9071	8	145750506	LRRC14	8	145743376	1.61E-50	5.20E-41	YES
rs11601413	11	18421683	LDHC	11	18433854	1.99E-49	1.47E-42	YES
kgp3928819	11	18439927	LDHC	11	18433854	1.83E-53	2.96E-44	YES
kgp6997008	11	18431265	LDHC	11	18433854	2.34E-48	3.79E-39	YES
rs3135006	6	32667119	HLA-DQB1	6	32636160	1.89E-39	1.85E-36	YES
kgp10470881	19	9264401	OR7D2	19	9296279	2.05E-83	1.10E-80	YES
kgp10544853	19	9250984	OR7D2	19	9296279	6.79E-83	3.91E-80	YES

Table S5.1: Top 30 eQTL ranked by BBER-adjusted FDRs. * "kgp" identifiers designates SNPs from 1000 Genome Projects pilot

MarkerName	P.value	PMID	disease
rs10134365	2.50E-05	19300482	Chronic obstructive pulmonary disease
rs10769813	8.54E-05	19300482	Chronic obstructive pulmonary disease
rs1080879	7.69E-07	19300482	Chronic obstructive pulmonary disease
rs10861369	9.39E-05	19300482	Chronic obstructive pulmonary disease
rs10926189	9.95E-05	19300482	Chronic obstructive pulmonary disease
rs10942957	3.15E-05	19300482	Chronic obstructive pulmonary disease
rs10989511	7.27E-05	19300482	Chronic obstructive pulmonary disease
rs11210569	4.40E-05	19300482	Chronic obstructive pulmonary disease
rs11219732	4.47E-06	19300482	Chronic obstructive pulmonary disease
rs11588172	9.02E-06	19300482	Chronic obstructive pulmonary disease
rs11603831	2.83E-06	19300482	Chronic obstructive pulmonary disease
rs11673894	9.71E-05	19300482	Chronic obstructive pulmonary disease
rs11719713	7.22E-05	19300482	Chronic obstructive pulmonary disease
rs12193019	2.27E-05	19300482	Chronic obstructive pulmonary disease
rs12421122	3.52E-05	19300482	Chronic obstructive pulmonary disease
rs12495172	2.79E-05	19300482	Chronic obstructive pulmonary disease
rs12576370	7.59E-05	19300482	Chronic obstructive pulmonary disease
rs12577504	8.16E-05	19300482	Chronic obstructive pulmonary disease
rs12662524	8.11E-05	19300482	Chronic obstructive pulmonary disease
rs12681897	4.84E-05	19300482	Chronic obstructive pulmonary disease
rs12762979	4.59E-05	19300482	Chronic obstructive pulmonary disease
rs12772513	6.08E-05	19300482	Chronic obstructive pulmonary disease
rs12796185	1.79E-05	19300482	Chronic obstructive pulmonary disease
rs13061634	2.76E-05	19300482	Chronic obstructive pulmonary disease
rs1402769	4.46E-05	19300482	Chronic obstructive pulmonary disease
rs1432534	8.75E-05	19300482	Chronic obstructive pulmonary disease
rs1551133	2.38E-05	19300482	Chronic obstructive pulmonary disease
rs1575208	1.46E-05	19300482	Chronic obstructive pulmonary disease
rs158990	5.04E-05	19300482	Chronic obstructive pulmonary disease
rs1622472	6.71E-05	19300482	Chronic obstructive pulmonary disease
rs163574	6.77E-05	19300482	Chronic obstructive pulmonary disease
rs16903825	2.35E-05	19300482	Chronic obstructive pulmonary disease
rs16943236	7.68E-05	19300482	Chronic obstructive pulmonary disease
rs16980016	7.95E-05	19300482	Chronic obstructive pulmonary disease
rs17099345	6.93E-05	19300482	Chronic obstructive pulmonary disease
rs17161994	9.46E-05	19300482	Chronic obstructive pulmonary disease
rs17310770	6.65E-05	19300482	Chronic obstructive pulmonary disease
rs1738899	6.09E-05	19300482	Chronic obstructive pulmonary disease
rs1828591	1.00E-07	19300482	Chronic obstructive pulmonary disease
rs1996020	4.86E-05	19300482	Chronic obstructive pulmonary disease
rs200303	7.96E-05	19300482	Chronic obstructive pulmonary disease
rs2080798	1.63E-05	19300482	Chronic obstructive pulmonary disease

rs2080799	6.74E-05	19300482	Chronic obstructive pulmonary disease
rs215864	3.30E-05	19300482	Chronic obstructive pulmonary disease
rs2248540	6.99E-06	19300482	Chronic obstructive pulmonary disease
rs2253023	2.46E-05	19300482	Chronic obstructive pulmonary disease
rs2269640	7.54E-05	19300482	Chronic obstructive pulmonary disease
rs239349	3.09E-05	19300482	Chronic obstructive pulmonary disease
rs25796	6.43E-05	19300482	Chronic obstructive pulmonary disease
rs2589183	4.50E-05	19300482	Chronic obstructive pulmonary disease
rs26566	7.81E-05	19300482	Chronic obstructive pulmonary disease
rs2898879	1.35E-05	19300482	Chronic obstructive pulmonary disease
rs2912263	3.59E-05	19300482	Chronic obstructive pulmonary disease
rs2963162	1.54E-05	19300482	Chronic obstructive pulmonary disease
rs30539	3.38E-06	19300482	Chronic obstructive pulmonary disease
rs312495	7.60E-05	19300482	Chronic obstructive pulmonary disease
rs32447	6.46E-05	19300482	Chronic obstructive pulmonary disease
rs32466	8.31E-07	19300482	Chronic obstructive pulmonary disease
rs3767943	9.18E-05	19300482	Chronic obstructive pulmonary disease
rs3847554	9.03E-05	19300482	Chronic obstructive pulmonary disease
rs3901366	1.21E-05	19300482	Chronic obstructive pulmonary disease
rs4148777	9.07E-05	19300482	Chronic obstructive pulmonary disease
rs4432437	4.22E-05	19300482	Chronic obstructive pulmonary disease
rs4588237	7.66E-05	19300482	Chronic obstructive pulmonary disease
rs4712564	8.79E-05	19300482	Chronic obstructive pulmonary disease
rs4739642	6.99E-05	19300482	Chronic obstructive pulmonary disease
rs4754595	4.90E-05	19300482	Chronic obstructive pulmonary disease
rs4842213	3.77E-05	19300482	Chronic obstructive pulmonary disease
rs4951563	1.73E-05	19300482	Chronic obstructive pulmonary disease
rs4974153	9.55E-05	19300482	Chronic obstructive pulmonary disease
rs6100861	7.16E-05	19300482	Chronic obstructive pulmonary disease
rs6426962	1.94E-05	19300482	Chronic obstructive pulmonary disease
rs6542127	9.25E-05	19300482	Chronic obstructive pulmonary disease
rs6706895	7.90E-05	19300482	Chronic obstructive pulmonary disease
rs6720264	6.69E-05	19300482	Chronic obstructive pulmonary disease
rs6790122	2.93E-05	19300482	Chronic obstructive pulmonary disease
rs6823107	3.62E-05	19300482	Chronic obstructive pulmonary disease
rs6852830	6.66E-05	19300482	Chronic obstructive pulmonary disease
rs6944889	7.21E-05	19300482	Chronic obstructive pulmonary disease
rs6973373	3.32E-05	19300482	Chronic obstructive pulmonary disease
rs7105963	9.90E-05	19300482	Chronic obstructive pulmonary disease
rs7233241	7.76E-05	19300482	Chronic obstructive pulmonary disease
rs729319	3.36E-05	19300482	Chronic obstructive pulmonary disease
rs732285	4.86E-05	19300482	Chronic obstructive pulmonary disease
rs7341022	9.97E-08	19300482	Chronic obstructive pulmonary disease

rs735243	2.02E-07	19300482	Chronic obstructive pulmonary disease
rs7522756	1.44E-05	19300482	Chronic obstructive pulmonary disease
rs7524799	6.74E-06	19300482	Chronic obstructive pulmonary disease
rs7529406	6.13E-05	19300482	Chronic obstructive pulmonary disease
rs7727670	8.38E-08	19300482	Chronic obstructive pulmonary disease
rs7775523	5.12E-05	19300482	Chronic obstructive pulmonary disease
rs7988287	4.69E-05	19300482	Chronic obstructive pulmonary disease
rs8009673	9.54E-05	19300482	Chronic obstructive pulmonary disease
rs8022070	8.10E-05	19300482	Chronic obstructive pulmonary disease
rs8034191	1.00E-10	19300482	Chronic obstructive pulmonary disease
rs9350301	8.93E-05	19300482	Chronic obstructive pulmonary disease
rs935381	7.47E-05	19300482	Chronic obstructive pulmonary disease
rs9482826	5.44E-05	19300482	Chronic obstructive pulmonary disease
rs9686327	9.84E-08	19300482	Chronic obstructive pulmonary disease
rs9788469	1.43E-05	19300482	Chronic obstructive pulmonary disease
rs9862661	1.57E-05	19300482	Chronic obstructive pulmonary disease
rs9978132	7.22E-05	19300482	Chronic obstructive pulmonary disease
rs1051730	2.14E-05	20173748	Chronic obstructive pulmonary disease
rs1062980	3.42E-05	20173748	Chronic obstructive pulmonary disease
rs13118928	9.30E-08	20173748	Chronic obstructive pulmonary disease
rs13180	2.00E-08	20173748	Chronic obstructive pulmonary disease
rs1903003	7.74E-08	20173748	Chronic obstructive pulmonary disease
rs2869967	1.80E-07	20173748	Chronic obstructive pulmonary disease
rs7671167	1.00E-11	20173748	Chronic obstructive pulmonary disease
rs10065677	7.60E-06	20709820	Emphysema-related traits
rs1012036	5.00E-06	20709820	Emphysema-related traits
rs10844154	4.80E-08	20709820	Emphysema-related traits
rs10900114	8.40E-06	20709820	Emphysema-related traits
rs1348350	3.10E-06	20709820	Emphysema-related traits
rs161981	6.10E-07	20709820	Emphysema-related traits
rs2129590	4.80E-06	20709820	Emphysema-related traits
rs2448276	5.20E-06	20709820	Emphysema-related traits
rs261869	8.10E-06	20709820	Emphysema-related traits
rs2999399	6.00E-06	20709820	Emphysema-related traits
rs326633	1.10E-06	20709820	Emphysema-related traits
rs341672	8.50E-06	20709820	Emphysema-related traits
rs4905179	9.80E-06	20709820	Emphysema-related traits
rs641525	5.00E-07	20709820	Emphysema-related traits
rs7905537	8.00E-07	20709820	Emphysema-related traits
rs7911712	3.80E-06	20709820	Emphysema-related traits
rs7977375	4.00E-06	20709820	Emphysema-related traits
rs8016091	4.10E-06	20709820	Emphysema-related traits
rs9292394	2.00E-06	20709820	Emphysema-related traits

rs8050136	4.00E-08	21037115	Body mass in chronic obstructive pulmonary disease
rs10928927	3.00E-07	21685187	Chronic obstructive pulmonary disease
rs114216682	7.00E-08	21685187	Chronic obstructive pulmonary disease
rs117607728	4.00E-07	21685187	Chronic obstructive pulmonary disease
rs28675338	1.00E-07	21685187	Chronic obstructive pulmonary disease
rs55645543	5.00E-07	21685187	Chronic obstructive pulmonary disease
rs56238310	1.00E-07	21685187	Chronic obstructive pulmonary disease
rs73717741	3.00E-07	21685187	Chronic obstructive pulmonary disease
rs76351433	2.00E-07	21685187	Chronic obstructive pulmonary disease
rs76884941	1.00E-07	21685187	Chronic obstructive pulmonary disease
rs77155169	2.00E-07	21685187	Chronic obstructive pulmonary disease
rs9296092	6.00E-07	21685187	Chronic obstructive pulmonary disease
rs9394152	7.00E-08	21685187	Chronic obstructive pulmonary disease
rs11858836	1.00E-06	22080838	Chronic obstructive pulmonary disease
rs13141641	3.00E-07	22080838	Chronic obstructive pulmonary disease
rs1964516	2.00E-09	22080838	Chronic obstructive pulmonary disease
rs7937	3.00E-09	22080838	Chronic obstructive pulmonary disease
rs2074488	2.00E-10	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs2077224	2.00E-14	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs1923539	5.00E-09	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs9951925	2.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs13181561	6.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs954820	4.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs6667220	8.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs3751143	4.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs12149070	8.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs17832777	6.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs7953249	1.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs652520	2.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs7078012	5.00E-09	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs3741240	1.00E-26	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs2463822	1.00E-10	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs17157266	1.00E-09	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs7929679	7.00E-09	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs8048576	9.00E-13	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs728616	2.00E-12	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs3851050	1.00E-11	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs12220777	7.00E-11	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs6585424	1.00E-10	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs9266629	4.00E-10	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs1265093	6.00E-09	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs3130559	8.00E-09	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs4508864	6.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers

rs2823743	1.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs1124480	9.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs903614	3.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs7006821	5.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs10007052	1.00E-07	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs7147624	5.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers
rs4468361	8.00E-06	23144326	Chronic obstructive pulmonary disease-related biomarkers

Table S5.2: List of COPD-associated SNPs from GWASdb

SNP	Chr	Position	Gene	Chr	tss	P-value	BBER-adj FDR	cis
kgp6093110	1	37958339	ODF1	8	103563800	4.30E-08	6.61E-05	No
rs7544462	1	37962756	ODF1	8	103563800	4.30E-08	6.61E-05	No
rs2579644	2	120219940	PPFIA2	12	82153332	2.03E-06	8.33E-04	No
rs9812727	3	55990726	DMRTC2	19	42349086	9.28E-08	3.99E-05	No
rs4622984	4	155452233	SOSTDC1	7	16570205	4.54E-07	1.87E-04	No
kgp4707314	4	145496941	C12orf61	12	62997214	2.34E-06	9.60E-04	No
kgp10989644	5	14990914	CLDN15	7	100882101	1.01E-06	4.13E-04	No
kgp10447643	6	31095942	C1orf105	1	172389828	1.54E-06	6.31E-04	No
rs3815087	6	31093587	PSORS1C1	6	31082527	9.29E-17	1.08E-09	Yes
rs9263699	6	31093699	PSORS1C1	6	31082527	1.30E-16	1.52E-09	Yes
rs3130557	6	31094703	VARS2	6	30876019	7.15E-07	2.94E-04	Yes
kgp6354565	6	31095210	PSORS1C1	6	31082527	1.70E-15	1.98E-08	Yes
rs9263715	6	31095801	PSORS1C1	6	31082527	9.29E-17	1.08E-09	Yes
kgp9957759	6	31096561	PSORS1C1	6	31082527	1.08E-13	1.26E-06	Yes
rs9263719	6	31096575	PSORS1C1	6	31082527	4.02E-13	4.69E-06	Yes
rs2074488	6	31240431	POU5F1	6	31148508	2.70E-08	1.27E-05	Yes
kgp6834213	7	3487754	TMEM128	4	4249950	2.92E-07	1.69E-04	No
rs1012036	7	52472450	ATP6V0A1	17	40610862	4.04E-07	1.66E-04	No
kgp141150	8	2734822	CCDC70	13	52436117	9.11E-08	4.89E-05	No
rs11987190	8	2738093	CCDC70	13	52436117	9.11E-08	4.89E-05	No
rs4948917	10	45339090	TMEM63A	1	226070069	3.55E-07	3.48E-04	No
rs4948917	10	45339090	LPA	6	161087407	1.79E-07	8.82E-05	No
rs4948917	10	45339090	ZBTB5	9	37465396	6.35E-08	3.13E-05	No
rs4948917	10	45339090	C12orf29	12	88427623	3.65E-08	1.80E-05	No
rs4948917	10	45339090	GCSH	16	81130008	1.15E-07	5.65E-05	No
rs1830888	10	45339904	ATP5A1	18	43684300	4.41E-07	3.25E-04	No
rs4948917	10	45339090	ZNF333	19	14800613	9.89E-08	4.87E-05	No
rs10769813	11	7642031	CYB5R2	11	7698453	7.87E-11	5.79E-08	Yes
kgp9157600	14	104808367	KIAA0196	8	126104082	1.66E-06	7.03E-04	No
kgp9157600	14	104808367	OR7E24	19	9361606	3.83E-07	2.05E-04	No
kgp11359642	15	78813334	CHRNA5	15	78857862	8.25E-15	1.12E-07	Yes
rs12906951	15	78825562	CHRNA5	15	78857862	6.28E-18	8.53E-11	Yes
rs12915366	15	78831753	CHRNA5	15	78857862	2.93E-18	3.98E-11	Yes
rs12916483	15	78832397	CHRNA5	15	78857862	6.28E-18	8.53E-11	Yes
rs3813571	15	78832792	CHRNA5	15	78857862	6.28E-18	8.53E-11	Yes
kgp8874223	15	78833612	CHRNA5	15	78857862	6.28E-18	8.53E-11	Yes
rs4886571	15	78833758	CHRNA5	15	78857862	6.28E-18	8.53E-11	Yes
rs4887062	15	78837801	CHRNA5	15	78857862	1.71E-16	2.32E-09	Yes
rs8053	15	78841220	CHRNA5	15	78857862	5.35E-18	7.26E-11	Yes
rs12907966	15	78843051	CHRNA5	15	78857862	8.04E-18	1.09E-10	Yes
kgp5396992	15	78844386	CHRNA5	15	78857862	8.04E-18	1.09E-10	Yes
rs3743077	15	78894896	CHRNA5	15	78857862	7.65E-11	1.25E-04	Yes

(continued)
Table S5.3: eQTL which are also GWAS SNPs. * "kgp" identifiers designates SNPs from 1000 Genome Projects pilot

SNP	Chr	Position	Gene	Chr	TSS	P-value	BBER-adj FDR	cis
kgp386011	4	145659198	TIMM17A	1	201924619	5.36E-07	2.20E-04	No
kgp3206674	4	145706070	TGM4	3	44916100	1.51E-06	7.32E-04	No
kgp1594216	4	145778772	TGM4	3	44916100	1.12E-06	4.60E-04	No
kgp386011	4	145659198	MYL3	3	46923659	1.53E-06	6.28E-04	No
kgp386011	4	145659198	COPB2	3	139108574	4.45E-07	1.83E-04	No
kgp386011	4	145659198	MED12L	3	150803484	7.97E-07	3.75E-04	No
rs6537319	4	145761160	GABRA6	5	160974069	2.24E-07	9.20E-05	No
kgp386011	4	145659198	SP8	7	20826505	1.03E-06	4.80E-04	No
kgp1594216	4	145778772	FGF19	11	69519410	5.26E-07	2.16E-04	No
kgp3206674	4	145706070	FGF19	11	69519410	7.45E-07	3.33E-04	No

Table S5.4: eQTL of HHIP SNPs, an important COPD GWAS region. * "kgp" identifiers designates SNPs from 1000 Genome Projects pilot